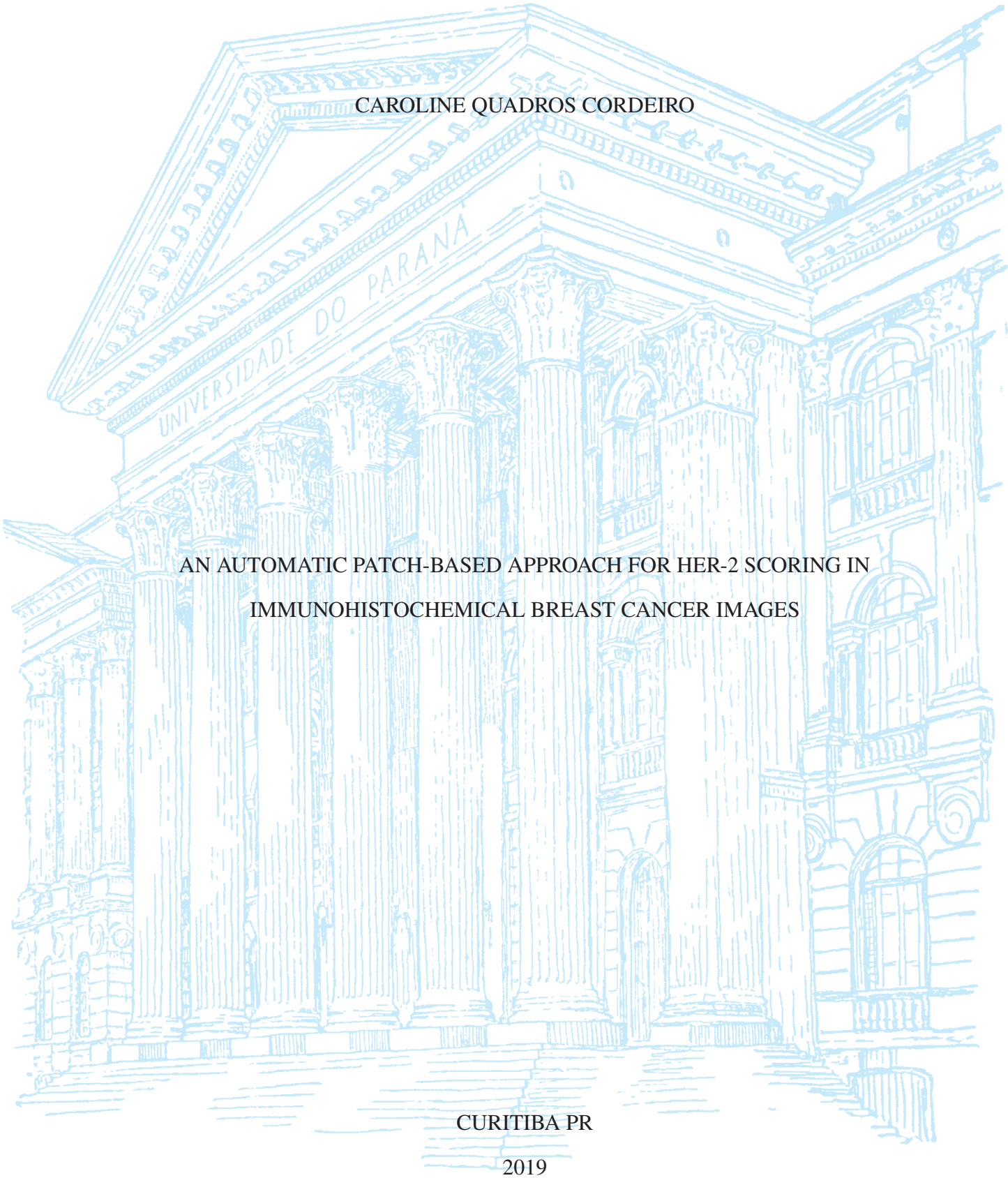
UNIVERSIDADE FEDERAL DO PARANÁ

CAROLINE QUADROS CORDEIRO



AN AUTOMATIC PATCH-BASED APPROACH FOR HER-2 SCORING IN IMMUNOHISTOCHEMICAL BREAST CANCER IMAGES

CURITIBA PR

2019

CAROLINE QUADROS CORDEIRO

AN AUTOMATIC PATCH-BASED APPROACH FOR HER-2 SCORING IN IMMUNOHISTOCHEMICAL BREAST CANCER IMAGES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Lucas Ferrari de Oliveira.

Coorientador: Sergio Ossamu Ioshii.

CURITIBA PR

2019

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

C794a

Cordeiro, Caroline Quadros
An automatic patch-based approach for Her-2 scoring in immunohistochemical breast cancer images [recurso eletrônico] / Caroline Quadros Cordeiro. – Curitiba, 2019.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2019.

Orientador: Lucas Ferrari de Oliveira. Coorientador: Sergio Ossamu Loshii.

1. Mamas – Câncer. 2. Processamento de Imagens. 3. Sistemas de reconhecimento de padrões. I. Universidade Federal do Paraná. II. Oliveira, Lucas Ferrari de. III. Loshii, Sergio Ossamu. IV. Título.

CDD: 004

Bibliotecária: Vanusa Maciel CRB- 9/1928

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA - 40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **CAROLINE QUADROS CORDEIRO** intitulada: **An Automatic Patch-Based Approach for HER-2 Scoring In Immunohistochemical Breast Cancer Images**, sob orientação do Prof. Dr. LUCAS FERRARI DE OLIVEIRA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 07 de Outubro de 2019.

LUCAS FERRARI DE OLIVEIRA
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

LUIZ EDUARDO SOARES DE OLIVEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

SERGIO OSSAMU IOSHII
Coorientador - Avaliador Externo (DEPARTAMENTO DE PATOLOGIA BáSICA)

ANDRÉ GUSTAVO HOCHULI
Avaliador Externo (PONTIFICA UNIVERSIDADE CATÓLICA DO PARANÁ)

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

À todos aqueles que se interessam pela ciência e pelo conhecimento. Que buscam inovar e aplicar a computação para beneficar à sociedade, principalmente no âmbito da saúde.

AGRADECIMENTOS

Devo resumir os agradecimentos reconhecendo a mim mesma como uma mulher de muita sorte, minha lista aqui seria enorme. Por isso, deixo a cargo de todos meus amigos, familiares, colegas e professores reconhecem a si próprios como seres pelos quais tenho enorme gratidão.

RESUMO

O Câncer de Mama (CaM / *Breast Cancer (BC))* é o câncer mais comum entre as mulheres em todo o mundo. O Instituto Nacional de Câncer (INCA) estima 59.700 novos casos de CaM em 2018 no Brasil. Em 2012, a taxa de mortalidade do CaM aumentou 14%. No Brasil, a 'Lei dos 60 Dias', estabeleceu que o primeiro tratamento oncológico no Sistema Único de Saúde (SUS), deve começar dentro de um prazo máximo de 60 dias a partir da assinatura do laudo patológico. Como um exame fundamental para definir a terapia adequada para pacientes com CaM, e um importante prognosticador, a análise do *Human Epidermal growth factor Receptor-type 2* (HER-2) é uma prática rotineira em laboratórios de patologia. Aproximadamente 20-25% dos CaMs são HER-2 positivos, portanto podem ser tratados com trastuzumab. Tal análise é visual e manual, é uma tarefa altamente especializada, demorada, extremamente dependente da experiência dos patologistas e diretamente influenciada por fatores como fadiga e diminuição da atenção. Assim, é propenso a erros, levando à variabilidade inter-patologistas nos resultados dos testes, o que pode afetar a precisão do diagnóstico. Para garantir a precisão do diagnóstico, patologistas e oncologistas rotineiramente solicitam segunda opinião. No entanto, uma segunda opinião nem sempre é facilmente acessível e pode levar várias semanas. Recentes avanços na Patologia Digital e no poder de processamento dos computadores permitiram o desenvolvimento de Software de Análise de Imagem Digital para ajudar nesta questão. Bases de dados com anotações são importantes para avaliar as soluções propostas. Portanto, uma das contribuições deste trabalho é introduzir uma nova bases de dados pública de *Whole Slide Images* (WSIs). O presente estudo pretende propor um algoritmo automático para pontuação do HER-2. Com base em diversas características, para desenvolver um sistema totalmente automatizado e livre de segmentação.

Palavras-chave: Pontuação HER-2, Câncer de Mama, Processamento de Imagens, Reconhecimento de Padrões, Aprendizado de Máquina, Patologia Digital, Whole Slide Image

ABSTRACT

Breast Cancer (BC) is the most common cancer among women worldwide. National Institute of Cancer (INCA) estimates 59,700 new BC cases in 2018 in Brazil. In 2012, the mortality rate for BC increased by 14%. In Brazil, the '60 Day Law', established the first oncological treatment in the Unified Health System (SUS) should start within a maximum period of 60 days from the signature of the pathological report. As a fundamental exam to defining the appropriate therapy for patients with BC, and an important prognosticator, the analysis of Human Epidermal growth factor Receptor-type 2 (HER-2) is a routine practice in pathology laboratories. Approximately 20-25% of BCs are HER-2 positive, thus they can be treated with trastuzumab. Such analysis is visual and manual, it is a highly specialized task, time-consuming, extremely dependent on the experience of the pathologists and directly influenced by factors such as fatigue and decrease of attention. Thus it is error-prone, leading to inter-pathologists variability in the tests results, which can affect diagnostic accuracy. To ensure diagnostic accuracy, pathologists and oncologists routinely request a second opinion. However, a second opinion is not always easily accessible and can take several weeks. Recent advances in Digital Pathology and processing power of computers allowed the development of Software of Digital Image Analysis to help in this issue. Annotated datasets are important to evaluate proposed solutions. Therefore, one of the contributions of this work is to introduce a new public dataset of Whole Slide Image (WSI). The present study intends to propose an automatic algorithm for HER-2 scoring. Based on different several types of features, to develop a fully automated and segmentation free system.

Keywords: HER-2 score, Breast Cancer, Image Processing, Pattern Recognition, Machine Learning, Digital Pathology, Whole Slide Image

LISTA DE FIGURAS

LISTA DE TABELAS

List of Acronyms

BC Breast Cancer

BG Background

CAD Computer-Aided Diagnosis

CART Classification and Regression Trees

CEP17 Centromere 17

CNN Convolutional Neural Networks

DAB Diaminobenzidine

DIA Digital Image Analysis

DT Decision Tree

EDM Euclidean Distance Map

FC Fully-connected

FCM Fuzzy c-means clustering

FG Foreground

FISH Fluorescence in Situ Hybridization

GEC Gastroesophageal Cancer

GLCM Grey Level Co-occurrence Matrices

GT Ground Truth

HE Haematoxylin-Eosin

HED Haematoxylin-Eosin-Diaminobenzidine

HER-2 Human Epidermal growth factor Receptor-type 2

HSV Hue, Saturation, Value

HVS Human Visual System

IARC International Agency for Research on Cancer

IHC Immunohistochemistry

INCA National Institute of Cancer

KNN K-Nearest Neighbor

LBP Local Binary Pattern

MCD Minimum Cluster Distance

MLP Multilayer Perceptron

PFTAS Parameter-free Threshold Adjacency Statistic

RF Random Forest

RGB Red-Green-Blue

ROI Region of Interest

SUS Unified Health System

SVM Support Vector Machine

TAS Threshold Adjacency Statistic

TLSTM Trapezoidal Long Short-Term Memory

UK United Kingdom

WHO World Health Organization

WSI Whole Slide Image

SUMÁRIO

1 INTRODUCTION

Cancer is a disease characterized by uncontrolled growth and spread of cells. The World Health Organization (WHO) estimates 27 million new cancers worldwide by 2030 and 17 million deaths from the disease. In the last decade, the incidence of cancer has grown by 20% in the world [46].

Breast Cancer (BC) is the second most common tumor worldwide. In the US one out of eight women is affected by BC during their lifetime [13]. In Brazil, BC is the most common among women, affecting almost 60,000 patients in 2014 [29]. The Brazilian National Institute of Cancer (INCA) estimates 59,700 new BC cases in 2018 [30]. According to the International Agency for Research on Cancer (IARC), while cancer mortality rate increased by 8% in 2012, the mortality rate of BC was 14% in the same period [16].

In order to increase patients survival, WHO has emphasized the recommendations for adoption of policies that favor early diagnosis, coupled with appropriate treatment in a timely manner [6]. In Brazil, the '60 Day Law' established that the first oncological treatment, in its Unified Health System (SUS), should start within a maximum period of 60 days from the signature of the pathological report or in a short term, according to the need of the case registered in the patient's medical record [79]. In 2017, the Brazilian Federal Government published in the Federal Official Gazette a decision to provide treatment with trastuzumab [11].

In BC patients, the amplification of the *Human Epidermal growth factor Receptor-type 2 (HER-2)* gene is an individual prognosticator and a predictive marker of response to targeted treatment with trastuzumab and adjuvant chemotherapy [63]. Various studies have asserted that the trastuzumab anti-HER2 monoclonal antibody is efficient in the treatment of the different BC stages [53].

For HER-2 score determination, immunohistochemical tests are performed. The HER-2 test indicates whether this protein is carrying some role in the development of BC, since with many HER-2 receptors, the cells receive many signals to grow and split. The amount of HER-2 is scored as 0, 1+, 2+ or 3+. If the score is 0 or 1+, it is called "HER-2 negative"; if the score is 2+, then it is called "borderline"; and a 3+ score is called "HER-2 positive"[35]. Approximately 20-25% of BCs are HER-2 positive [80]. The early use of adjuvant trastuzumab therapy in patients HER2-positive reduces the risk of mortality [53].

HER-2 scoring still has a visual and manual analysis of histological tissues as a standard method. Such method is strongly dependent on the expertise and experience of histopathologists and has the disadvantages of being time-consuming and non-replicable [2]. Some HER-2 tests may present different results, indicating the existence of variations within and between specialists observation [35].

There are other factors which can affect results and lead to divergences, such as color variations in Immunohistochemistry (IHC). This variations can be caused by ischemic time, tissue fixative and fixation time, tissue processing, the efficiency of epitope retrieval, selection of antibody or its clone and detection system [74].

In this context, considering the relevance of cancer in public health and the benefit of a correct treatment, as well as the opportunity to provide supporting tools for pathologists; we propose to investigate an automated method for BC HER-2 scoring in Whole Slide Image (WSI).

1.1 MOTIVATION

Image Processing in scanned slides has received significant attention due to the widespread use of slide scanners (digital microscopes) and the application of computational algorithms for image analysis, which has contributed to the proper use of biomarkers that can be used in stratified medicine [23]. Advances in software development and improved computing capacity have also led to an increase in interest in Digital Pathology [50].

In routine clinical practice, the histopathological analysis is based on the opinion of pathologists, which visually analyze the tissue under the microscope. Such visual inspection is a highly specialized task, error-prone, time-consuming and directly influenced by factors such as fatigue and low attention [31].

Significant diagnostic variability has been reported between pathologists and it is inferred that 4% of negative cases and 18% of positive cases are misdiagnosed. In particular, scoring variability has been shown to be important for cases that show heterogeneous HER-2 expression within the tumor cell population [76], due to this heterogeneity that should be assessed and taken into account when determining treatment options [8].

According to some studies, applications of automated image analysis for HER-2 protein expression, instead of manual assessment, decrease the need for supplementary Fluorescence in Situ Hybridization (FISH) testing by 68% [27]. Which can be a time and financial advantage, as shown in Table 1.1.

Tabela 1.1: Comparison of Immunohistochemistry and Fluorescence In Situ Hybridization as Screening Tools for HER-2 in Breast Cancer. Adapted from [80].

	FISH	**IHC**
Failure rate (%)	5.0	0.08
Procedure time (mean/std)	36h/30min	4h/13s
Interpreting time by the pathologist (mean/std)	7min/2.5min	45s/13s
Mean direct reagent cost for laboratory ($)	140	10

Often, pathologists and oncologists seek a second opinion to ensure the accuracy of the diagnosis. However, a second opinion can demand a long time. In the last decade, various algorithms have been created for computer-assisted HER-2 classification. Nonetheless, most systems are commercial and dependent on specific and financially costly materials, or they are algorithms in which the coefficient of agreement with pathologists is not considered sufficient for application in diagnostic practice [2, 73, 62, 41, 21].

1.2 OBJECTIVES

Aiming to propose an HER-2 scoring approach in BC, our objectives with this work are:

- To develop an automatic HER-2 scoring system, avoiding segmentation and manual intervention;
- To evaluate the potential utility of Computer-Aided Diagnosis (CAD) to facilitate clinical decision-making;
- To evaluate different computational methods to scoring HER-2 in BC;
- To create a dataset of histopathological images in which IHC tests for HER-2 were performed, including variability in preparation of the material.

1.3 CONTRIBUTIONS

The contributions of this work are following described.

- **Development of a new HER-2 scoring system.** A new fully automatic and without segmentation HER-2 scoring system was develop in this work. Hopefully, the system can work as a second opinion for medical experts, supporting mainly inexperienced pathologists, reducing workload and making treatment decisions more accurate. Besides, decreasing the number of cases requiring subsequent FISH, which is more expensive than IHC.

- **Public dataset.** This work introduces a new public dataset of histopathological images including WSIs, which were tested for HER-2. We provide the Ground Truth (GT) according to the clinical reports issued and reviewed by three experienced pathologists. This dataset is a collaboration with the *Hospital Erasto Gaertner* and has been approved by the ethical committee on 28/03/2018 (CAAE: 84415418.5.0000.0098, approval number 2.568.281 2.568.281).

1.4 DOCUMENT STRUCTURE

The presented document is composed of 6 chapters. In the current chapter, we discussed the relevance of BC in worldwide public health. We also highlighted the importance of early treatment, which requires a correct and urgent decision. Motivated by these topics, we presented the HER-2 scoring problem, our objectives, and contributions with this work.

In Chapter 2 we offer a basic background for a proper understanding of this work, the required knowledge to analyze the results and the proposed approach.

A review of the state-of-the-art for HER-2 scoring in BC is given in Chapter 3.

Chapter 4 contains our proposed methodology and the description of our new dataset named HistoBC-HER2.

Chapter 5 shows the results. The experiments carried out to date were performed in the Warwick dataset [50].

Finally, Chapter 6 is the conclusion of the work.

2 FUNDAMENTAL BACKGROUND

For a proper understanding of the present work, this chapter provides some fundamental background knowledge about the type of image, color space, features, and classifiers used in this work.

2.1 HER-2 SCORING PROBLEM

HER-2 is a gene that can play a role in the development of BC. Some genes and the proteins can influence how a BC behaves and how it might respond to a specific treatment. Cancer cells from a tissue sample might be tested to see which genes are normal and abnormal. The proteins, as HER-2, may also be tested [45].

The HER-2 IHC scoring method is a semi-quantitative system based on the intensity of the reaction product and percentage of membrane positive cells, giving a score range of 0–3+ (Figure 2.1). Samples scoring 3+ are regarded as unequivocally positive, and those scoring 0/1+ as negative. Borderline scores (2+) are regarded as equivocal and mandate further assessment using FISH [52].

The United Kingdom (UK)'s guideline recommendations for HER-2 testing is detailed in Table 2.1.

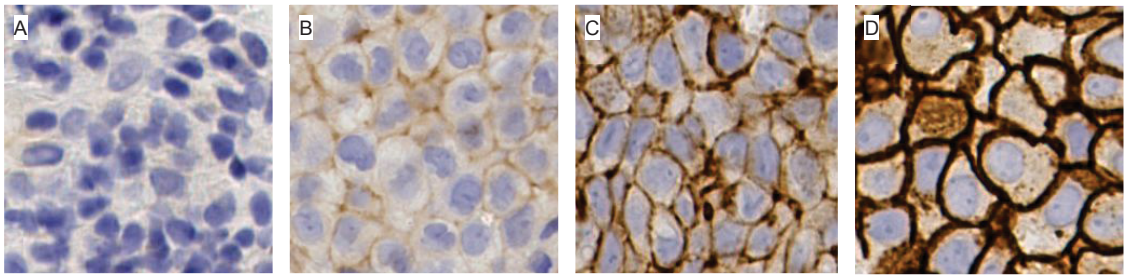


Figura 2.1: Example of HER-2 scores.
(A) 0. (B) 1+. (C) 2+. (D) 3+.
Source: The author.

Tabela 2.1: Recommended automated HER-2 scoring criteria for IHC-stained BC tissue slides.

Score	Staining Pattern	Classification
0	No membrane staining or incomplete membrane staining in less than 10% of invasive tumor cells	Negative
1+	Faint/barely perceptible membrane staining or weak incomplete membrane staining in more than 10% of tumor cells	Negative
2+	A weak to moderate complete membrane staining is observed in more than 10% of tumour cells or strong complete membrane staining in less than 10% of tumor cells	Borderline
3+	A strong (intense and uniform) complete membrane staining is observed in more than 10% of the invasive tumor cells	Positive

2.2 TYPE OF IMAGE

In this work, we tested two WSIs datasets - the University of Warwick [50] and our own dataset named HistoBC-HER2.

WSI is a type of image for scanned slides, which have a pyramidal structure to enable optimized viewing across multiple magnification levels, providing a high-resolution overview of the entire slide.

Rather than one small microscopic field, a WSI consists of the creation of a single, high magnification digital image of an entire microscopic slide. In summary, an automated microscope scans an entire slide at one or more resolutions, then combines consecutive small images into a single large image — normally some gigabytes in size.

Typically, at x40 magnification, the images have a resolution of approximately 0.25 microns per pixel. At this resolution, a slide region of size 15mm x 15mm could correspond to 60,000 x 60,000 pixels [42].

One example of different magnification levels viewing is shown in Figure 2.2. A region is selected in the slide and amplified by the magnification of x10, x20 and x40.

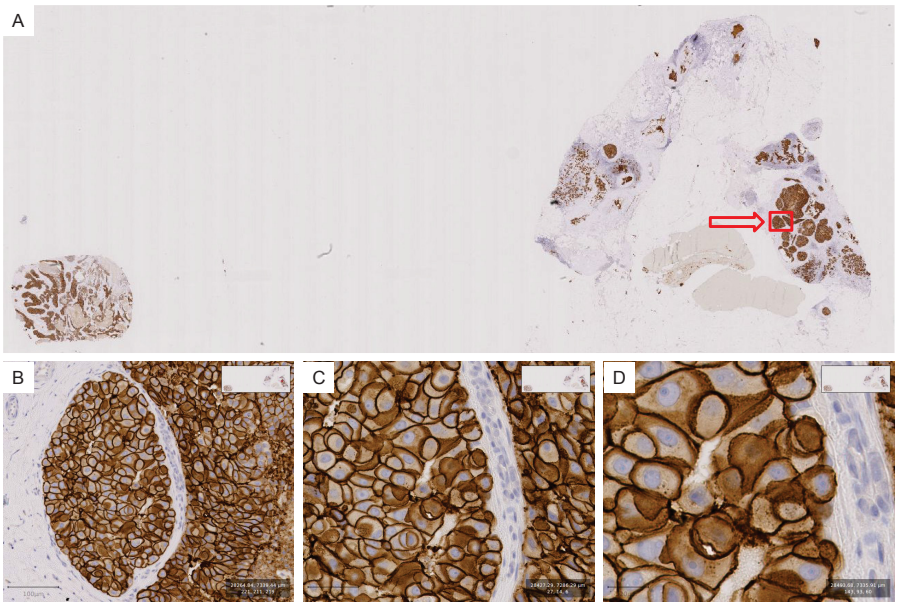


Figura 2.2: Example of magnifications views. (A) Entire slide. (B) Selected part in x10. (C) Selected part in x20. (D) Selected part in x40.
Source: The author.

2.3 COLOR SPACE

Color is the way the Human Visual System (HVS) measures a part of the electromagnetic spectrum, approximately between 300 and 830 nm. A color space is a notation by which we can specify colors in the human perception of the visible electromagnetic spectrum [72].

2.3.1 RGB

Generally, electronic systems described color in Red-Green-Blue (RGB) color model. It is an additive color system in which a color is a composite of three primary colors: red, green and blue.

The RGB system can be represented by a cube in a cartesian coordinate system using values within a 0–1 range, the color black is represented in this cube's corner (0, 0, 0), and red, green, and blue are along the x, y, z-axis, as shown in Figure 2.3.

The diagonal of the cube from the origin is a line representing equal parts of red, green, and blue along its length and therefore, as the distance from the origin increases, the color of such points on the diagonal goes from black to white [39].

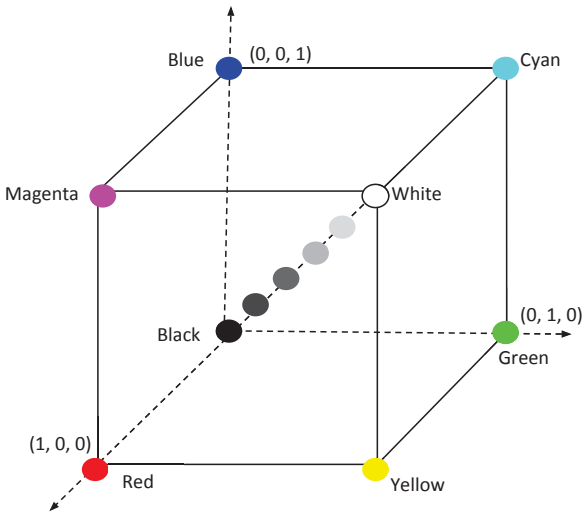


Figura 2.3: RGB color system representation.
Source: The author.

A color corresponds to a point in the RGB color space, with coordinates (R, G, B), and all realizable colors must correspond to points that lie within the cube.

This scheme corresponds well to the actual physical devices used to produce colors with computers. However, the perceptions of human observers are better described by other color systems which correspond more directly to the subjective human sensation of color. One of the simplest and most widely used of these alternate systems is the Hue, Saturation, Value (HSV) color system [39].

2.3.2 HSV

Unfortunately, the RGB color model is not well suited for describing colors in terms that are practical for human interpretation. HSV color model is more similar to the way we perceive colors, describing it by its hue, saturation, and brightness.

The HSV color space is formed by Hue (H), Saturation (S) and Value (V). In this model, the hue is a color attribute that describes a pure color. Saturation gives a measure of the degree to which a pure color is diluted by white light. And Value is actually the brightness, which is a subjective descriptor that is practically impossible to measure. It embodies the achromatic notion of intensity and is one of the key factors in describing the color sensation.

HSV model can be represented as a cone, as shown in Figure 2.4. In this illustration, the component H can be described as an angle, with a domain between $[0, 2\pi]$. Saturation can be understood as the radial distance (from the center) of the cone, assuming values between [0, 1]. And V component can be represented as the vertical axis of the cone and has values belonging to the interval [0, 1].

Consequently, HSV is an ideal tool for developing image processing algorithms based on color descriptions that are natural and intuitive to humans [18].

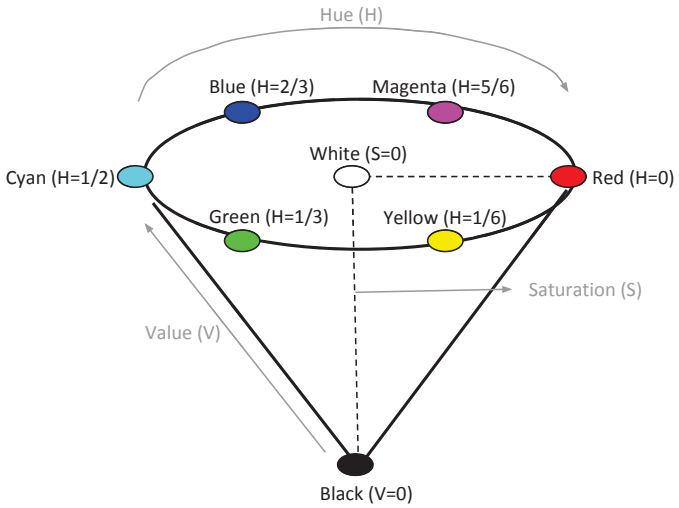


Figura 2.4: HSV color system representation.
Source: The author.

2.3.3 HED

Haematoxylin-Eosin-Diaminobenzidine (HED) is a color space in which Scikit-image [75] allows representing IHC stained images.

One of the most common stainings is Haematoxylin-Eosin (HE). Basically, hematoxylin stains the cell nuclei in blue and eosin stains cytoplasm in magenta-red [57]. Diaminobenzidine (DAB) has been widely used as a chromogen for revealing protein expression by means of IHC together with HE for tissue counterstaining, as a positive result, a tissue area is stained in brown.

Ruifrok and Johnston [57] have proposed a color deconvolution algorithm that allows the separate presentation of stain components. A result is shown in Figure 2.5.

2.4 FEATURE EXTRACTION

Extracting features is one of the most critical tasks in pattern recognition and it generally affects the end result more than the choice of classification algorithm [54, 15]. A feature extraction algorithm aims to represent an image MxN in a single vector of d-dimensional characteristics, where d is smaller than MxN. In this section, we briefly describe the different operators used for feature extraction.

2.4.1 Texture Extractors

Although texture concept is not precisely defined, an image can be characterized in terms of its visual appearance. Texture properties are coefficients that describe the image and, they are obtained by exploring the spatial relationships underlying the gray level distribution.

The idea of using texture to describe histopathological images is to have a simple but powerful representation. Texture representation allows avoiding explicit segmentation of cells structures, such as nuclei and membranes.

2.4.1.1 Local Binary Pattern (LBP)

Ojala *et al.* (1996) [44] proposed a descriptor with a concept that the texture of an image can be decomposed into a set of small textural units.

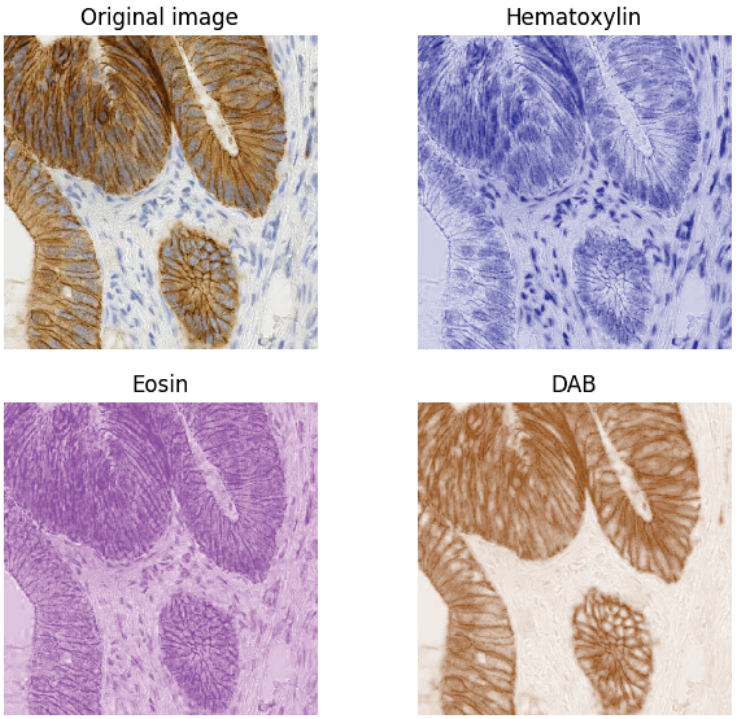


Figura 2.5: HED color space representation. Image from [75].

Each pixel receives a label LBP according to the gray levels of its neighbor pixels. This neighborhood is defined by a number of P pixels equally distributed over a circumference of radius R centered on the analyzed pixel. LBP label is obtained by thresholding the neighborhood with the central pixel as the threshold. The binary pixels are then multiplied by power values of 2, corresponding to the position of the pixel. The descriptor $LBP_{P,R}$ is the histogram of these 2^P labels.

Since a neighbor pixel initially in the right side of the central pixel if rotated creates a different LBP label, it is necessary to remove the effect of rotation, assign a unique LBP label to each rotation invariant local binary pattern. $LBP^{ri}_{P,R}$ is a LBP rotation invariant, which generates a histogram of 36 bins for $P = 8$, as illustrated in Figure 2.6. This variation was proposed in [43], the main idea is to rotate the binary pattern clockwise P times and obtain the minimum LBP label.

Ojala *et al.* (2002) [43] noticed certain local binary patterns, called 'uniform', can contain more information than others, thus they are fundamental properties of texture. A pattern is considered to be uniform if it has at most 2 bits transitions. This variation is called $LBP^{u2}_{P,R}$ and its descriptor is a histogram of 59 bins, 58 for uniform patterns and one for non-uniform. Figure 2.7 shown the 58 uniform patters.

Similarly, the rotation invariant a uniform version $LBP^{riu2}_{P,R}$ considers all the P possible rotations for the uniform patters. Thus, the descriptor is a histogram of 10 positions, 9 for the uniform patterns and 1 for non-uniform.

2.4.1.2 Parameter-free Threshold Adjacency Statistic (PFTAS)

PFTAS, proposed by Coelho *et al.* (2010) [9], is similar to Threshold Adjacency Statistic (TAS), a method presented by Hamilton *et al.* (2007) [22]. The idea is to threshold the

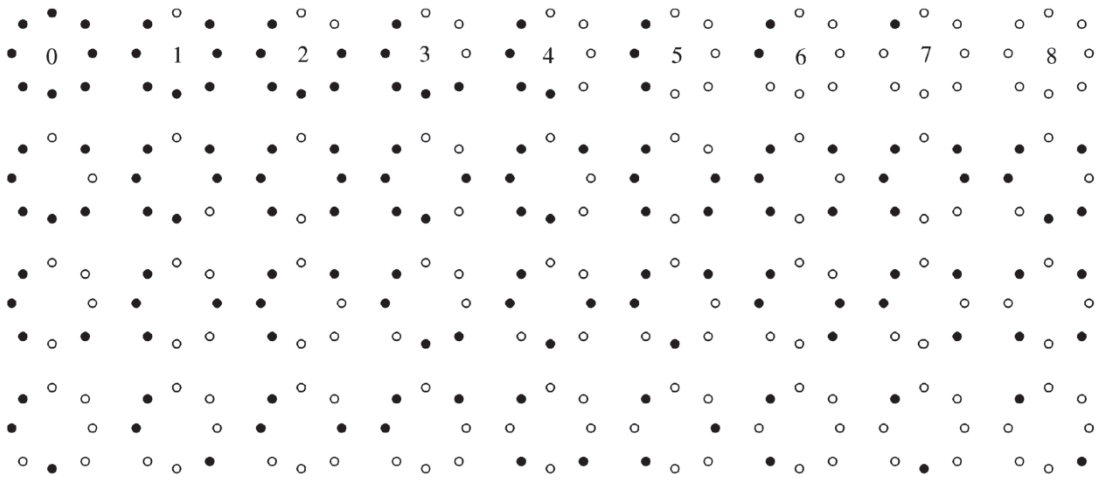


Figura 2.6: The 36 unique invariant binary patterns that can occur in the circularly symmetric neighbor set of $LBP^{ri}_{8,R}$. Image from [43].

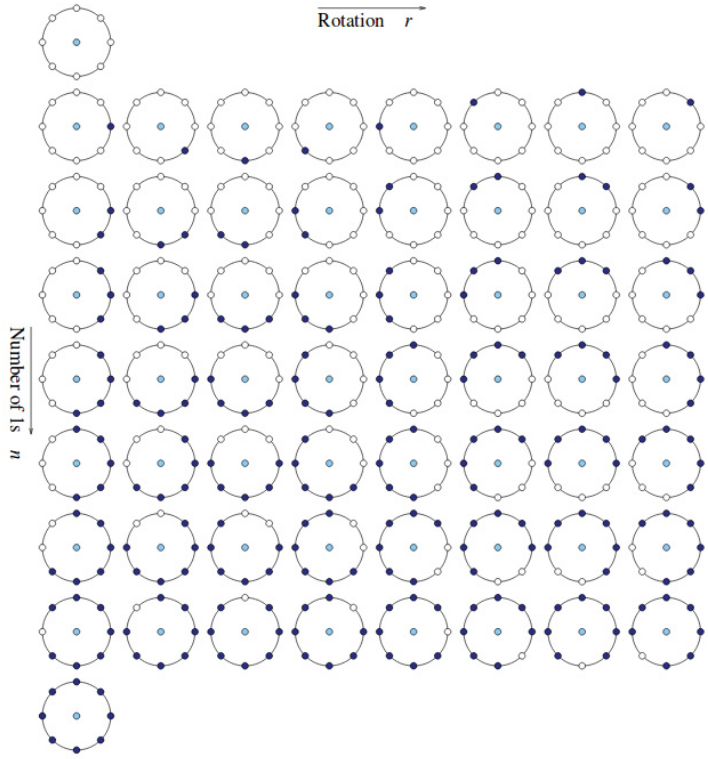


Figura 2.7: The 58 different uniform patterns in (8,R) neighborhood. Image from [1].

image using three different thresholds, $[\mu - \sigma, \mu + \sigma]$, $[\mu - \sigma, 255]$ and $[\mu, 255]$, where μ is the threshold defined by Otsu algorithm - the step which makes the algorithm parameter-free. For each white pixel in these binary images, nine statistics about the number of white adjacent pixels are computed, according to Figure 2.8.

Examples of having zero to eight white neighbors are given in Figure 2.8. The first threshold statistic is then the number of white pixels with zero white neighbors - P(0), the second is the number of white pixels with one white neighbor - P(1), and so on up to eight. These

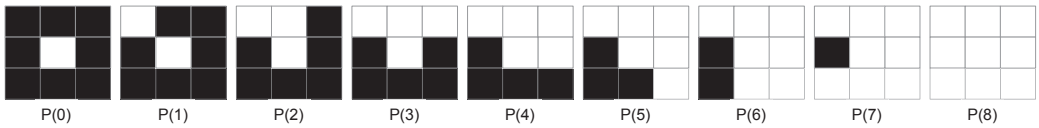


Figura 2.8: The nine threshold adjacency statistics calculated in PFTAS. Image adapted from [22].

nine statistics are then normalized by dividing each by the total number of white pixels in the thresholded image.

2.4.1.3 Grey Level Co-occurrence Matrices (GLCM)

GLCM is a method of texture analysis proposed by Haralick in 1973 [25]. It consists of a statistical method that analyzes the distribution of co-occurring pixel gray-scale values in the image.

The GLCM matrix G is a symmetric matrix of size $N_g \times N_g$, where N_g is the number of grey levels in the image. The (i, j) position in the matrix G contains information on how many times the i^{th} and j^{th} pixel values occur in a given distance d and angle θ.

In total, 14 measures were proposed by Haralick *et al.* to be extracted from GLCM [25] [24]. However, some works in the literature affirm that there are redundancies between the information of these 14 measures, thus, only a subset of four metrics was used in this work: contrast, dissimilarity, homogeneity, and energy. All extracted from the matrices generated by distances 1, 2 and 3 and angles 45º, 90º, 135º and 180º. The result is a vector of size 48.

2.5 CLASSIFIERS

Classification is a task that consists of evaluating the processed data, labeling them according to their characteristics. Supervised classification enables this task through a set of training data that will have its labels previously assigned by a specialist. Different classifiers were used to assess the HER-2 scoring problem. These classifiers are presented in the following sections.

2.5.1 Support Vector Machine (SVM)

Proposed by Vapnik and Cortes (1995) [10], Support Vector Machine (SVM) is a machine learning strategy initially used to classify into two groups. It tries to create a hyperplane to separate the classes. A hyperplane can be understood as a division of Euclidean space into two, each region containing data from a single class.

If the training set is arranged in a Euclidean space, then it is possible to find infinite hyperplanes satisfying the division of the regions into two distinct groups [70]. The best hyperplane for the separation is obtained by the following equation:

$$\vec{w} \cdot \vec{x_i} + b = 0$$

Where $\vec{w}$ is a vector of weight perpendicular to the hyperplane, $\vec{x}$ are the attributes of an example and b is a compensating factor that allows increasing the margin of the separation of hyperplanes. Both b and $\vec{w}$ are parameters adjusted during training [70].

The distribution of training data in Euclidean space allows us to determine to which region a new sample belongs to. If such a determination is made, it is possible to predict its class

c_j by assigning to the sample the same class of the other data in the region, provided one of the two conditions is satisfied [70]:

$$\begin{cases} \vec{w} \cdot \vec{x}_i + b \geq 1 & \text{se } c_j = 1; \\ \vec{w} \cdot \vec{x}_i + b \leq -1 & \text{se } c_j = -1. \end{cases}$$

Formally, SVM is able to differentiate between the two regions. This makes it conceptually a binary classifier. However, it is possible to use SVM as a multi-class classifier, just make use of one-against-one strategy [78].

As data is not always linearly separable, it can be necessary to map them to an N-dimensional space where this is possible. A kernel function allows performing this data mapping. Scikit-learn [47] provides the following kernels: 'linear', 'poly', 'rbf', 'sigmoid'.

2.5.2 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) consists of layers of neurons connected to each other by weights and output signals. Layers are named input, hidden and output. The input layer takes the inputs from the dataset and passes it to the network. The output layer makes the prediction about the input. An MLP may have one or more hidden layers between input and output layers. Figure 2.9 illustrates a model of MLP [17].

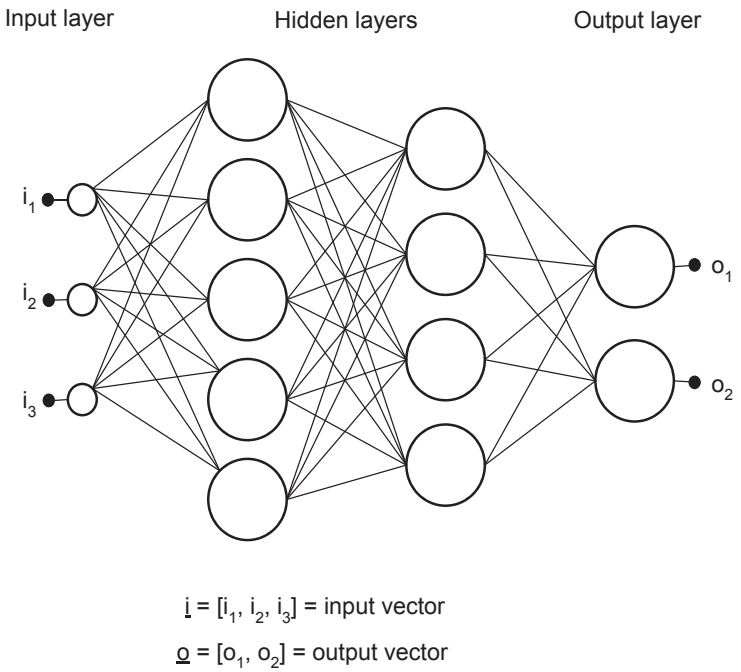


Figura 2.9: A multilayer perceptron with two hidden layers. Image adapted from [17].

The outputs of each of these neurons are computed and then passed along to the neurons in the next layer. This is called a feed-forward neural network [54].

The learning process occurs during training step, which requires a set of training data with inputs and their respective output vectors. The training adjusts the weights in the network by a repetition of training input, until matches input and output. In order to reduce the overall error in MLP, the difference between true output and predicted output is used to determine the adjustment in the weights in the network [17].

2.5.3 K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) classifier has the objective of finding within the training data, those relatively similar to the sample that will be predicted. The more similar K examples are known as K-nearest neighbors and are used to determine the class for the given data. This type of classifier represents each example as a point in a d-dimensional space, where d is the number of attributes. For a test example, its distance to the rest of the points in the training set is calculated.

Classifying an unknown sample using the K-Nearest Neighbor (KNN) algorithm consists of initially calculating the distance between the test sample and the training sample. There are several ways of performing this distance calculation, the simplest is the Euclidean distance [15]. Once the distances are obtained, the next step is to label the test data by assigning to it the same label as the majority among the K-nearest neighbors.

It is conceptually a fairly simple algorithm, and it is easy to implement since it does not process in the training step. On the other hand, as the training set increases, also the cost of the algorithm increases. Nearest neighbor classifiers make their predictions based on local information. Thus, they are susceptible to noise for small values of K.

2.5.4 Decision Tree (DT)

Decision Tree (DT) is a flowchart-like structure. The general idea is to create a model able to predict a class making decisions based on rules learned from training [54, 47].

The DT is composed of nodes that form a rooted tree. A root node has no incoming edges, meaning it is a directed tree. The intern nodes, or test node, have only one incoming edge and outgoing edges according to the tested attribute's value. In the case of numeric attributes, the condition refers to a range. The other nodes are called leaves or decision nodes [56]. Each internal node makes a decision about an attribute to determine the next node, until the leaf node which determines a class label [54].

The construction of a DT is usually top-down, at each step the variable in which best split training dataset is chosen for a node. The evaluation of the split can be done by different metrics, generally related to the homogeneity of the target variable within the subsets.

The *scikit-learn* package implements the Classification and Regression Trees (CART) an algorithm as its default DT class, which can use both categorical and continuous features. The metric evaluation used is Gini impurity and Variance reduction [56].

2.6 DEEP LEARNING

In supervised learning, we have examples of expected output, whereas in unsupervised learning some assumption is made to build the model. However, for the algorithm success is imperative to have good features, as they can well represent the data. To solve the problem of finding this representation, deep learning learns it from the data: it defines representations that are expressed in terms of other, simpler ones [49].

It can yield more nonlinear and more abstract representations, using an architecture formed by the composition of multiple levels of representation [3, 4]. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics [36].

Convolutional Neural Networks, also known as ConvNets, is probably the most well known Deep Learning model used to solve Computer Vision tasks, in particular, image classifica-

tion [49]. Convolutional Neural Networks (CNN), specifically, have achieved success in image classification problems, including medical image analysis [65].

CNN is a variant of MLP. They are composed of neurons that have learnable weights and biases. In summary, a CNN consists of multiple trainable stages stacked on top of each other, followed by a supervised classifier and sets of arrays named feature maps, which represent both input and output of each stage [38]. There are three main types of layers used to build CNN architectures: convolutional layer, pooling layer, and fully-connected layer. Normally, a complete CNN architecture is obtained by stacking several of these layers.

- **Convolutional:** The main constituent parts of CNNs are the convolutional layers, which are composed of a set of filters (or kernels) to be applied to the entire input. Each filter is nothing but a matrix of weights (or values) and each one is a feature to be learned. As a result of each filter is an affine transformation of the input. A filter is convolved around the input image, then the name of the layer [36].

- **Pooling:** Often applied after a few convolutional layers, the main objective of a pooling layer is to progressively reduce the spatial size of the representation [49]. In practice, the max-pooling function, which applies a window function to the input patch, and computes the maximum in that neighborhood, has shown better results [60]. However, the pooling units can perform other functions like L2-norm pooling or average pooling.

- **Fully-connected:** For classification, after many convolutional layers, it is common to include Fully-connected (FC) layers that work in a way similar to a hidden layer of an MLP. The feature maps of the last convolutional layer are vectorized and fed into FC layers followed by a softmax logistic regression layer [40, 49].

Architectures such as AlexNet [34], VGG [61], ResNet [26] and GoogLeNet [67] became very popular, used as subroutines to obtain representations that are then offered as input to other algorithms to solve different tasks [49].

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [61]. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. It was one of the famous model submitted to ILSVRC-2014 [58]. It improves AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another.

Since the pre-trained VGG-16 model has learned to extract features from images that can distinguish one image class from another, they have shown to achieve excellent performance even when applied to image recognition and classification datasets in other domains [20, 61].

Deeper than VGG net, Resnet has a depth of up to 152 layers, however still having lower complexity. This architecture won the 1st place on the ILSVRC-2015 [58] classification task addressed the degradation problem by introducing a deep residual learning framework. Instead of hoping every few stacked layers directly fit a desired underlying mapping, they explicitly let these layers fit a residual mapping [26].

2.7 EVALUATION METRICS

To evaluate the performance of a classifier, several metrics can be analyzed. The choice of such metrics depends on the problem. We chose to analyze the accuracy, precision, and recall.

Accuracy is basically the number of correct predictions divided by the total predictions, it means how much the classifier is in agreement with the pathologist. The formula is presented in Equation 2.1.

Precision is the agreement of the data class labels with those from a classifier. The formula is presented in Equation 2.2.

Recall is the effectiveness of a classifier to identify class labels. The formula is presented in Equation 2.3.

The literature suggests that for multi-class problems, a *per-class* evaluation may be more effective [64]. Thus, the formulas below are used to calculate the evaluation metrics in this work.

$$accuracy = \frac{\sum_i M_{ii}}{\sum_{i,j} M_{ij}} \tag{2.1}$$

$$precision_{\ i} = \frac{M_{ii}}{\sum_j M_{ji}} \tag{2.2}$$

$$recall_{\ i} = \frac{M_{ii}}{\sum_j M_{ij}} \tag{2.3}$$

3 RELATED WORKS

This chapter reviews the literature related to HER-2-assisted computer classification, especially CAD using histopathological image processing. The determination of HER-2 score in BC, both IHC and FISH are considered equally valuable methods. While IHC provides a measure of protein expression, FISH provides *HER-2* gene amplification [77, 41]. Reviewed works approached both, IHC and FISH images, each one of these approaches is described in separated sections below.

Although significantly increase of research in the field of digital pathology in benign/-malign classification, this state-of-the-art analysis will be limited to those works reporting HER-2 score on histopathological images.

3.1 FISH IMAGES

Relatively few works in literature have concentrated on FISH images. Mostly recent, Razavi (2017) [53] proposed a method to classify negative, equivocal and positive HER-2 status. His method used HSV channels to detect green (Centromere 17 (CEP17)) and red (HER-2) signals. Filters and morphological operations are applied. Following, a threshold and Euclidean Distance Map (EDM) are used as seeds for watershed segmentation. A rule based on the proportion of red and green signal determined the final HER-2 score.

Theodosiou *et al.* (2008) [71] presented an overall accuracy of 92.8%, with 100% in negative cases and 74.1% in positive cases. The authors developed a module to integrate into the Volumetric Image Processing, Analysis, and Visualization software package, EIKONA3D (Alpha Tec Ltd). FISH image evaluation is performed via the algorithm described in [51].

In [51] an algorithm to case classification by an overall image from a slice is described by Raimondo and Gavrielides (2005). Nuclei and spots are segmented. The spot segmentation consists of a top-hat filtering stage followed by template matching to separate real signals from noise. Nuclei segmentation includes a nonlinearity correction step, global thresholding to identify candidate regions and a geometric rule to distinguish between holes within a nucleus and holes between nuclei. Finally, the marked watershed transform is used to segment cell nuclei with markers detected as regional maxima of the distance transform.

3.2 IHC IMAGES

IHC is a process to detect protein expression localized in tissue cells using the anti-gen/antibody principle. The HER-2 test indicates whether this protein is carrying some role in the development of BC. The advantages of IHC testing include its wide availability, relatively low cost and easy preservation of stained slides [71]. Since the mean direct reagent cost by vendors of each FISH and IHC test was $140 and $10, respectively [80], an accurate diagnosis based on IHC tests and avoiding FISH is advantageous. Works present below are proposed solutions for HER-2 scoring in IHC images.

3.2.1 Classical Image Processing

The automatic imaging processing for cancer diagnosis has been explored as a topic of research since the 1970s [66]. Additionally, these techniques have also been tested for the purpose

of the HER-2 score in BC. In this section, we review the latest studies related to automatic image processing for scoring HER-2.

Mukundan (2017) [42] proposed an algorithm to HER-2 scoring contest [50] based on characteristic curves, he was the second-best points score. His idea was to threshold each channel in HSV images. Then, plot variations of the percentage of staining with color in some threshold range against saturation channel threshold values. The Validation dataset presented 88.46% of accuracy.

Patches in size 100x100 were selected by high entropy criteria in Tabakov *et al.* (2013) [69] work. In order to classify data from above 60 patients as 'recommended' and 'not recommended' for trastuzumab therapy, they proposed to use the Fuzzy Sugeno Integral, as an aggregation operator of an ensemble of fuzzy decision trees. Color values, structural factors, and texture information helped to build three different fuzzy DT. Accuracy on k-fold cross-validation was 83%.

Fuzzy c-means clustering (FCM) method was performed in HSV color space, to create a color palette which helped to create a normalized color histogram in [32], a work of Keller, Chen, and Gavrielides (2012). Linear regression on the training set was used to select the features (color histogram values) for the classification. A continuous score is given as output, providing an ordinal index of HER-2. In this work, Kendall τb was used as a measure of agreement. The range of Kendall τb is from -1 to 1, where 1 indicates that the readers are always concordant (perfect agreement), -1 indicates they are always discordant (perfect disagreement), and 0 indicates no agreement. They presented 0.72 Kendall τb.

Tuominen *et al.* (2012) [73] presented a publicly available web application for scoring HER-2. Their algorithm requires a blank field and a positive control image to normalize illumination and brown intensity in a test image. Membranes and nuclei are segmented using DAB deconvolution and filters like a median, following by threshold and morphological operation. They classified membranes as 'complete and strong staining' or 'incomplete or weak staining'. By combining these properties they scored HER-2 into 0/1+, 2+ or 3+. Analysis of the validation dataset showed a weighted kappa coefficient $\kappa_w = 0.80$. Nevertheless, as the normalization step should be provided by users, it can introduce errors and difficult the pathologist's tasks.

Masmoudi *et al.* (2009) [41] obtained an overall percentage of agreement in the order of 81%–83%. Their dataset was composed of 13 images for training and 64 for cross-validation. Only 1+, 2+, and 3+ classes were included in the dataset, 0 class was excluded. Their algorithm firstly does a color pixel classification to differentiate epithelial cell nuclei, epithelial cell membrane, and background. Two different linear regression were used, one to extract membrane pixels and another one to extract nuclei pixels. They implemented component analysis, watershed segmentation algorithm and hole-filling operation for nuclei segmentation. Features as membrane completeness and membrane staining intensity were extracted, with the objective of providing quantitative measures of HER-2 expression. The extracted features were used to classify each slide in a category of 1+, 2+ or 3+, using a Minimum Cluster Distance (MCD) classifier.

In work [21] DAB decomposition was applied followed by an Otsu's method and rotationally invariant bar filter to segment membrane stained with HER-2. In these experiments, Hall *et al.* (2008) reported three scoring features based on mean intensity. The first feature used to calculate a score is M_p, the mean intensity of stained membrane regions. The second feature is M_n, which is M_p normalized by the positive control tissue. The third feature M_a adds a coefficient d/N to M_n, where N is the total amount of pixels in the image and d is the number of DAB-stained pixels after preprocessing. Only 2+ cases were tested by leave-one-out validation. These cases were correctly classified 64% of the time, whereas manual scoring was only able to correctly classify 23% of the cases.

Based on built-in ImageJ program functions, Skaland *et al.* (2008) [62] created a macro to automatically segment membrane staining and measure the area of membrane fragment, mean and median staining intensity. Briefly, membrane staining was segmented by using the find maxima (segmented particles) function followed by converting to mask function. The mask was adjusted with two dilations and erosions. By a combination of color deconvolution, thresholding on the DAB component, and removing membrane fragments with less than 50 pixels, the image was cleaned up. Only 2+ and 3+ examples were evaluated and the correlation was 100%, probably overfitting occurred.

3.2.2 Deep Learning

Deep learning is a technique that has increased in importance over the last five years. It has quickly become the state of the art in computer vision [36]. In this section, we review the latest studies which assessed deep learning for scoring HER-2.

A deep learning framework for identifying, segmenting and classifying cell membranes and nuclei from HER-2 stained BC images was proposed by Saha and Chakraborty (2018) [59]. They created an approach using Trapezoidal Long Short-Term Memory (TLSTM). Authors performed various combinations of training and testing datasets and presented the average of following metrics, 96.64% precision, 96.79% recall, 96.71% F-score, 93.08% negative predictive value, 98.33% accuracy and a 6.84% false-positive rate. Images from the HER-2 scoring contest [50] were used in this study, however, they did not use the entire dataset.

Vahadane (2017) [74] proposed in his Ph.D. thesis an approach that was submitted to HER-2 scoring contest [50]. A deep CNN was trained to learn the patch the HER-2 score prediction and accumulate the patch scores to predict the HER-2 score for WSI through a criterion. In addition to the CNN learned features, handcrafted average control tissue information input to the first fully connected layer, in order to normalize the variations in IHC. The work did not present results, however, it was the first in the rank competition.

Vandenberghe *et al.* (2017) [76] were one of the first authors to investigate deep learning as a solution to the HER-2 automatic scoring problem. Classical machine learning techniques (SVM and Random Forest (RF)) were compared with ConvNets. They used color decomposition to separate the brown HER-2 staining and the blue hematoxylin staining. The tissue was then segmented into cells using the watershed algorithm. They selected representative regions for training. A total of 18 features for the classical approach were extracted, describing nucleus color, nucleus texture, nucleus morphology, HER-2 membrane staining intensity and proportion of positive HER-2 membrane staining. For deep learning, an image patch of 44x44 around each detected cell was extracted and directly used as input to the ConvNets model. As a result, they obtained 83% of accuracy and affirmed deep learning was not better than classical machine learning in distinguish between 2+ and 3+ classes.

Another approach using deep learning was developed by Rodner, Simon, and Denzler (2017) [55]. It was submitted to the HER-2 scoring contest [50]. Its architecture was based on AlexNet and pre-learned from ImageNet. They randomly crop several 227x227 patches at resolution level 1 (highest resolution is at level 0) within a 1024x1024 window around user click location. A single 227x227 image patch is passed to CNN and the activations layer was obtained at conv5. The activations, which was represented as a vector, were then transformed into a matrix by calculation Gramian matrix. Multi-class logistic regression is used to classify four classes. Due to the achievement of 100% accuracy using fine-tuning on a fixed partition training/validation, they assumed to probably have overfitting, as accuracy without a fine-tuning was about 75%.

Pitkäaho *et al.* (2016) [48] selected regions of interest in the lowest level resolution and then mapped to the highest. These regions were cut in 128x128 patches. They chose a CNN to classify each patch, the architecture is similar to the AlexNet architecture. Data was augmented by rotating each block three times (90-degree rotations), each resulting block was augmented through horizontal mirroring. Then, they create a rule for WSI classification, based on the patches percentage of each class. The problem with this approach is that they repeated images in testing and training, thus achieving 97.7% accuracy. Experiments were also assessed by HER-2 scoring contest images [50].

3.2.3 Commercial Systems

Currently, some HER-2 scoring systems are available on the market. Mostly are trained for a particular biomarker, consequently, they are not robust to laboratory variabilities. This section briefly presents some studies of these commercial HER-2 scoring systems.

A software named HER2-CONNECT™, which is a module of the Visiopharm Integrator System (VIS) platform was experimented by Koopman *et al.* (2018) [33]. This software was originally developed for BC, however, authors would like to evaluate the ability of the system to classify Gastroesophageal Cancer (GEC). Overall agreement between system and consensus manual scores was 76.5% for BC and 85.6% for GEC.

Authors of [27], in order to minimize the number of equivocal 2+ scores and the need for reflex FISH analysis, compared automated Digital Image Analysis (DIA) with manual reading. For this purpose, HER2-CONNECT™(Visiopharm) was studied. They conclude the manual assessment and DIA classification were identical, representing a concordance of 90.5% between the pathologist and the automated DIA algorithm.

The algorithm of HER2-CONNECT™is presented in the work presented by Brügmann *et al.* (2011) [7]. The first step in their algorithm is preprocessing images to identify brown pixels. Post-processing includes skeletonizing the membrane, eliminate small membrane fragments and merging membranes that are not perfectly connected. Membrane connectivity is calculated from the size distribution of all membrane fragments remaining after post-processing. By this connectivity, the algorithm calculates the HER-2 score. Their overall agreement was 92.1% (Cohen's Kappa: 0.859) in the training set and 92.3% (Cohen's Kappa: 0.864) in the validation set. The image analysis sensitivity was 99.2% and specificity 100% when correlated to FISH. Some parameters in this algorithm were tuning for specific staining and imaging conditions. Thus, in other laboratories, results can differ dramatically from those of the actual study, as affirmed by the authors.

SlidePath's Tissue IA system (Leica) was presented by Dobson *et al.* (2010) [14] and obtained 91% of concordance with the pathologist. The images were classified as 0/1+, 2+ or 3+. But the classes were not well distributed, as they have 183 negative images, 40 of 2+ and 52 of 3+. A comparison from this work with other commercial systems is presented in Table 3.1.

Tabela 3.1: Comparison of commercial system available.

Software	**Concordance**	**Quantitation base**
SlidePath	91%	Intensity, continuity
Aperio	86%	Intensity
BioImagene	81%	Morphology, intensity
Dako	75%	Intensity
Ventana	86%	Intensity

Other two commercial system, ACIS (Automated Cellular Imaging System III) (Dako) and ScanScope (Aperio) have the scoring reproducibility of HER-2 evaluated in [68]. The concordance between systems was 86.5%. ACIS presented 72.8% concordance with pathologist and Aperio presented 70.4%.

3.3 FINAL REMARKS

Several methods for the HER-2 score were previously described. Commercial systems were presented and also works with deep learning and classical image processing approaches. Table 3.2 summarizes the methods of several works using IHC test images.

We noticed a lack of works during the years 2014 and 2015. Complementary research would suggest works focused on benign/malignant classification in HE stained images and standardization of methods of visual analysis.

In the literature, most of the works use their own datasets which are not available for other researches. Evaluation metrics are not standardized, leading to difficult comparisons. Thus, there is a necessity of a public dataset and unified benchmarks.

Generally, a proposed solution includes segmentation, which is a problem because errors in this step might compromise the overall performance of the system. Moreover, manual intervention is a usual requirement, demanding time and attention to the pathologists.

Manual intervention is also a limitation of commercial systems, in the sense that they are trained for a particular biomarker set and need to be manually optimized. Such adjustments introduce subjective criteria and become sources of interlaboratory variability.

Since the commercial systems are expensive and have these limitations, like other solutions, new alternatives to the HER-2 scoring problem still have to be developed.

Tabela 3.2: Comparison among HER-2 scoring methods – Images from IHC tests.

[ref] (year)	Dataset	# Images	Proposed Solution	Manual Intervention	Segmentation	Result
[59] (2018)	Warwick [50]	Tr= 51 Te= 28	CNN, TLSTM	Yes	Yes	98.33% accuracy (in validation)
[74] (2017)	Warwick [50]	Tr= 52	CNN for patches and average control tissue Criteria for WSI	No	No	First in rank competition
[76] (2017)	Private	71	CNN Hand-crafted features + SVM Hand-crafted features + RF	No	Yes	68% accuracy 70% accuracy 83% accuracy
[55] (2017)	Warwick [50]	Tr= 52	Bilinear features from AlexNet ImageNet for patches characteristics Multi-class logistic regression classifier	Yes	No	75% accuracy without fine-tuning (in validation)
[48] (2016)	Warwick [50]	Tr= 52	AlexNet for patches Rule for WSI	Yes	No	97.7% accuracy (in validation)
[42] (2017)	Warwick [50]	Tr= 52	Characteristic curve on HSV channels	Yes	No	88.46% accuracy (in validation)
[69] (2013)	Private	60	Fuzzy Sugeno Integral Fuzzy DT	No	No	83% accuracy
[32] (2012)	Private	77	Color palette FCM	Yes	No	0.72 Kendall τb
[73] (2012)	Private	Tr= 220 Te= 144	Measurement of membrane pixels	Yes	Yes	k_w = 80 ASE= 0.08
[41] (2009)	Private	Tr= 13 Te= 64	Linear regression Membrane measurement MCD	Yes	Yes	81% accuracy

[21] (2008)	Private	99	DAB decomposition, threshold Membrane measurement	No	Yes	64% accuracy (only for 2+ cases)
[62] (2008)	Private	60	DAB decomposition, threshold, morphologic operations	Yes	Yes	100% accuracy Likely overfitting

4 PROPOSAL

We proposed a patch-based approach which divided the experiments into two levels: image and patient. Firstly, each WSI was split into patches of size 250x250. Then, the image level evaluates the classification of patches. And the patient level evaluates the classification of WSI, it means a final HER-2 score.

For the image level, a subset of patches was created and named *feat_tr*. Around 30 patches of each WSI were selected by a pathologist to compose it. Each patch received the same score as its correspondent WSI. Then, feature extractor and classifiers were experimented to analyze the ability to distinctive HER-2 score in small images, with less heterogeneity.

Thereby, we used the best algorithms and subset *feat_tr* as a training set to classify all patches in each WSI, creating then a histogram of the HER-2 score.

Subsequently, based on the analysis done in the image level, the HER-2 score is calculated at the patient level. This level uses the created histogram to classify the final result.

Figure 4.1 illustrates our algorithm pipeline for image and patient level. The green rectangle has a representation of the image level, where the patches are extracted from tissue and each patch is scored for HER-2. Meanwhile, the purple rectangle has an illustration of the patient level, where all the patches scores are combined to determine the patient's final HER-2 score.

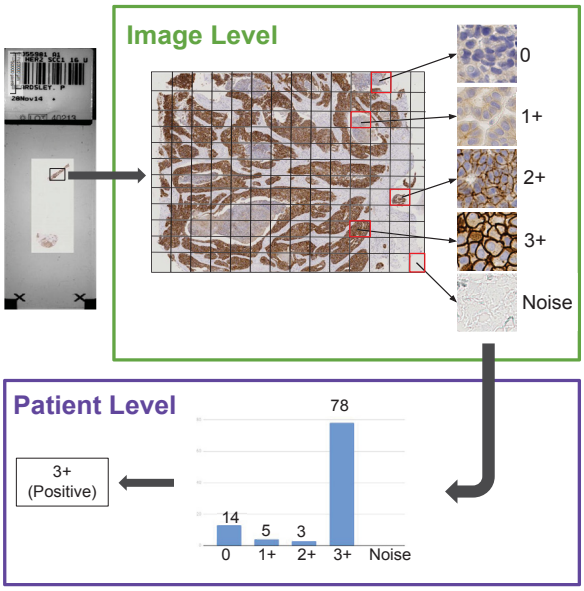


Figura 4.1: An illustration of our method.
Source: The author.

4.1 WARWICK'S DATASET

The proposed method was first developed based on the HER-2 image dataset of the Department of Computer Science, University of Warwick, UK [50]. The dataset entailed 172 WSI extracted from 86 cases of invasive breast carcinomas and included both the HE and HER-2 stained slides. Images stained with HE are used in the routine diagnostic practice of BC to identify tumor regions. Our approach only uses the HER-2 stained slides, being composed of 52 WSI for training and 34 for testing.

The histology slides for this contest were scanned on a Hamamatsu NanoZoomer C9600, enabling the image to be viewed from a ×4 to a ×40 magnification.

The authors of this dataset only provided GT for training images. It is required to submit the algorithm to evaluate testing images. The pathologist involved in our work reviewed training' GT. During this analysis, the pathologist suggested removing one image, due to it contained few cells, which would hinder a reliable analysis. The resulted WSI's classes distribution is presented in Table 4.1

Tabela 4.1: Classes distribution in Warwick's dataset.

Class	**#Training**	**#Testing**
0	9	4
1+	15	7
2+	14	13
3+	13	10
Total	51	34

4.2 OUR NEW DATASET - HISTOBC-HER2

For the purpose of evaluating the robustness of our algorithm, we create a new dataset. Due to this dataset includes staining color variation, it allows the evaluation under clinically realistic conditions. The HistoBC-HER2 dataset included HE and IHC stained slides.

Figure 4.2 shows an example of HistoBC-HER2 dataset image. The blue rectangle is the HE slide. The orange one is control tissue and the green one is the tested tissue, which was IHC stained.

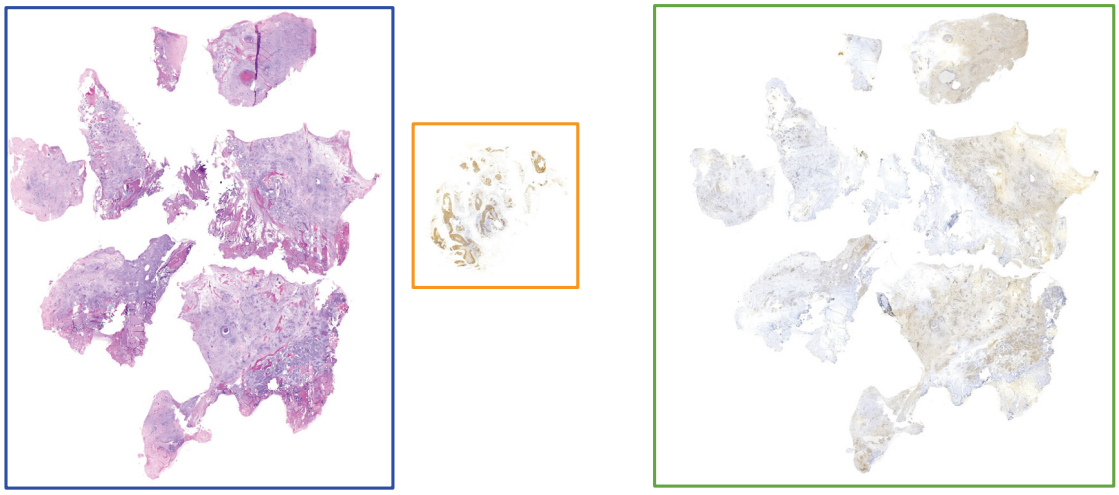


Figura 4.2: Example of HistoBC-HER2 images
Source: The author.

In order to meet the ethical aspects of research, as recommended by Resolution 466/2012 of the National Health Council[1], which provides guidelines and regulatory norms for research

[1]Resolução Nº 466 do Conselho Nacional de Saúde, de 12 de dezembro de 2012 (BR). Aprova as diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. Diário Oficial da União. 12 dez 2012.

involving human subjects, the project was submitted to the Ethics Committee on Research - *Hospital Erasto Gaertner*. Thus, according to CAAE 84415418.5.0000.0098 and approval number 2.568.281 2.568.281 of 28/03/2018, the ethical principles concerning this type of investigation were respected, and the project was duly approved by this committee.

The HistoBC-HER2 dataset contains 135 WSI of BC which were tested for HER-2. Images were collected through a clinical study from January 2012 to December 2012. Instead of collecting all tested images during this period, we collect only the amount balanced in all classes. All the data were anonymized. The HER-2 scores were reviewed by other two pathologists since we collect only the result in the electronic patient record and we want to avoid subjective criteria.

For the preparation of this dataset, it was necessary to rebuild the slides, as they were contaminated by fungi. For this, we had to dip the slides in chemicals for 24 hours, thus allowing the blade and coverslip to detach. After this a new coverslip was placed, allowing better visibility of the tissue under the microscope.

Subsequently, the slides were digitized in a Zeiss Axio Scan Z1 with the objective lens of x40. This scanner saves WSI in its own format, *.czi*. Since on that date, we did not find a support for this format, we also saved the *.jpg* images used to compose the WSI.

Finally, after obtaining the WSIs, a review was necessary to remove the images that contained the control tissue. As well as the review of the GT by two pathologists. Resulted WSI's classes distribution is presented in Table 4.2, where Rec means patient records and P1 e P2 are the pathologists.

Tabela 4.2: Classes distribution in HistoBC-HER2 dataset

Class	**#Rec**	**#P1**	**#P2**
0	32	61	1
1+	34	20	38
2+	35	20	63
3+	34	34	33

Table 4.2 shows how scores are different in each analysis. For example, P2 seems to be more biased to score 2+ and rarely scores 0. Details about each HER-2 score are presented in Tables 4.3, 4.4 and 4.5.

As HistoBC-HER2 has three annotations for each image, another classification was included for result analysis. We classified images in hard, medium and easy to classify, according to the number of agreement among all annotations. To clarify, if all annotations agreed, the image is considered easy to classify. An image with two agreements has a medium classification, and with zero agreements it a hard image to classify. This analysis resulted in 47 easy, 69 medium and 19 hard images to classify.

Figure 4.3 shows three views of exam number 40, which is an example of a hard image to classify. This exam was scored 0, 1+ and 2+. In this figure is possible to see the percent of marked cells and also how completed they are.

4.3 PATCH-APPROACH

Due to training a gigapixel resolution WSI is currently computationally impossible we decided to use a patch-approach. As suggested by Hou *et al.* (2016) [28], training a patch-level classifier on image patches will perform better than or similar to an image-level classifier.

Tabela 4.3: HER-2 scores in HistoBC-HER2 dataset - Easy images

Easy images											
Exam	Rec	P1	P2	Exam	Rec	P1	P2	Exam	Rec	P1	P2
001	3	3	3	224	2	2	2	513	3	3	3
020	0	0	0	233	3	3	3	518	3	3	3
056	3	3	3	238	3	3	3	519	2	2	2
058	3	3	3	239	3	3	3	541	3	3	3
095	3	3	3	251	2	2	2	631	3	3	3
097	3	3	3	276	3	3	3	649	2	2	2
103	3	3	3	304	2	2	2	661	3	3	3
121	2	2	2	334	3	3	3	662	3	3	3
136	3	3	3	338	3	3	3	667	1	1	1
141	2	2	2	358	3	3	3	670	3	3	3
149	3	3	3	398	3	3	3	679	2	2	2
169	3	3	3	447	3	3	3	695	2	2	2
178	3	3	3	470	3	3	3	748	2	2	2
203	3	3	3	476	1	1	1	864	2	2	2
207	3	3	3	479	3	3	3	865	2	2	2
				491	0	0	0	887	2	2	2

Tabela 4.4: HER-2 scores in HistoBC-HER2 dataset - Medium images

Medium images											
Exam	Rec	P1	P2	Exam	Rec	P1	P2	Exam	Rec	P1	P2
003	1	0	1	202	0	0	1	442	1	2	2
008	1	0	1	216	1	2	2	448	2	1	2
015	0	0	1	225	1	0	1	451	1	0	1
016	0	0	1	232	1	1	2	457	1	0	1
018	0	0	2	266	1	1	2	458	2	1	2
024	2	0	2	307	2	0	2	460	2	1	2
043	1	1	2	329	1	0	1	477	0	0	1
049	1	1	2	331	0	0	1	478	0	0	2
054	1	0	1	335	0	0	1	490	1	0	1
063	0	0	1	349	2	3	3	498	1	1	2
073	0	0	1	355	0	2	2	499	0	0	2
074	0	0	1	360	0	0	1	520	2	0	2
077	0	0	1	363	0	0	2	534	0	0	1
102	2	1	2	366	0	0	1	588	2	1	2
161	0	0	1	377	2	0	2	625	1	0	1
163	2	1	2	378	2	0	2	630	2	3	2
166	0	0	1	380	0	0	1	693	1	1	2
168	3	2	2	382	0	0	1	699	2	0	2
170	0	0	1	385	1	2	2	785	2	1	2
171	2	1	2	386	0	0	1	823	2	1	2
182	0	0	1	391	0	0	1	859	2	1	1
191	3	2	3	400	0	0	2	875	2	1	2
195	0	0	1	441	1	2	2	899	2	1	2

Each WSI was cropped in patches of size of 250x250 pixels at ×40 magnification. WSIs of Warwick's dataset are in *.ndpi* format, which is supported by the OpenSlide library [19].

Tabela 4.5: HER-2 scores in HistoBC-HER2 dataset - Hard images

Hard images											
Exam	Rec	P1	P2	Exam	Rec	P1	P2	Exam	Rec	P1	P2
017	0	3	1	184	1	0	2	379	1	0	2
022	1	0	2	214	1	0	2	444	1	0	2
051	1	0	2	218	1	0	2	459	1	0	2
053	2	0	3	220	1	0	2	480	1	0	2
062	3	0	2	273	1	0	2	638	3	0	1
096	2	0	1	325	0	3	2	668	1	0	2
								707	1	0	2

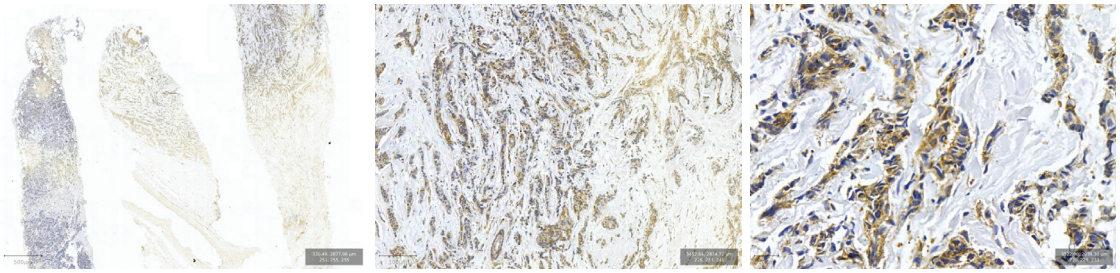


Figura 4.3: An example of a hard image to classify.
Source: The author.

The patches with more tissue information were automatically selected by analyzing their histograms. We deal with this patches selection in two steps: (1) separate Foreground (FG) and Background (BG); and (2) select in FG those with more tissue information.

1. It analyzes the patch contains pixels with a value above a threshold, if yes, consider as FG.

2. It analyzes if FG patches already selected contain enough information. Each bin in the histogram is multiplied by a factor and then we sum up. We consider as relevant the ones with the result above a certain threshold.

The parameters were empirically decided. This patch selection is important to reduce the number of images to be processed, as we show in Chapter 5. Figure 4.4 has examples of this approach. In the first line are examples of patches filtered in step 1, with separates background and foreground. The second and the third lines are examples of patches consider as foreground, although only the third line has examples of patches considered relevant.

4.4 FEATURE EXTRACTION

The purpose of this step is to state the features which best describe the patches. To achieve this purpose, we create a subset named *feat_tr*. This subset is important due to it is used to evaluate the feature vectors and will later be used as a training set for the classification of the total set of patches of each WSI.

Assisted by a pathologist, we selected around 30 patches out of each WSI to compose *feat_tr*. As a criterion for this selection, we asked the pathologist to choose the patches that best represent the class of their respective WSI. This amount of 30 was decided to balance the relevant ones among total patches of each WSI. Figure 4.5 shown some examples of patches from *feat_tr*.

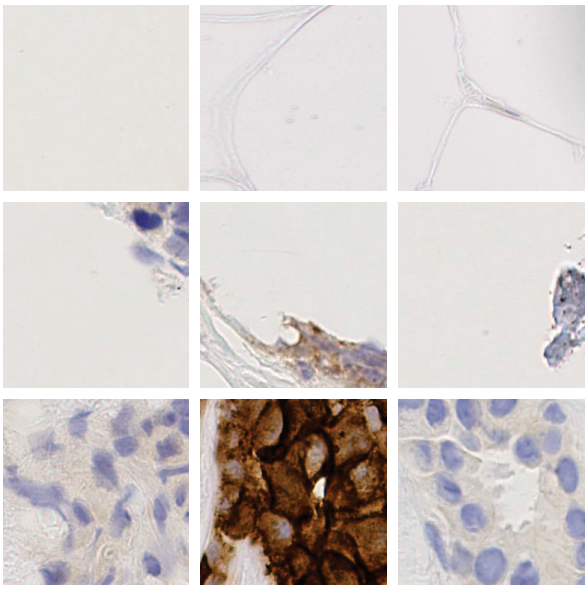


Figura 4.4: Patches Selection. Each line has examples of patches classified as background, foreground with few information and foreground with relevant information.
Source: The author.

The subset *feat_tr* can be interpreted as images of Region of Interest (ROI) of the slides. Due to these patches have characteristic tissue information, it is possible to compare with a selected region by a pathologist in a non-automatic approach with manual intervention.

The features were evaluated using leave-one-patient-out validation in the subset *feat_tr*.

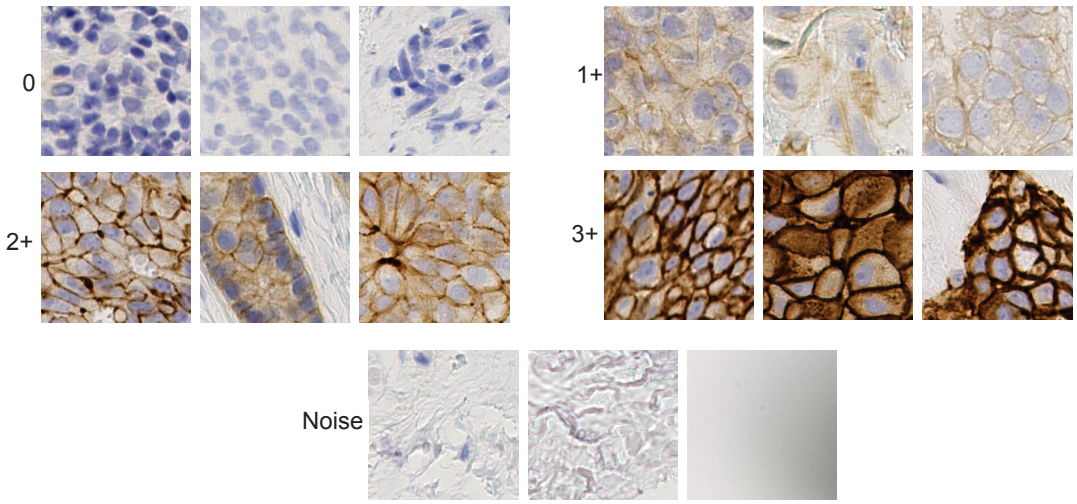


Figura 4.5: Examples of patches from *feat_tr* - those selected to evaluate features.
Source: The author.

As some patches remained after patch-selections steps and do not have tissue information, we included another class, the 'noise' one. A total of 364 representative patches of this class was selected. The resulted distribution of subset *feat_tr* is presented in Table 4.6.

We experimented LBP and PFTAS as texture descriptors, different variations of the LBP algorithm and changing parameters. These descriptors were applied in the original and deconvoluted images. We also try color features as some statistical features in HSV and RGB spaces. Furthermore, we use CNN to extract features.

Tabela 4.6: Classes distribution in *feat_tr*.

Class	#Patches
0	277
1+	387
2+	419
3+	420
Noise	364
Total	1867

In order to summarize results present in Chapter 5 we use the following notation for the features:

- **HSV:** Histogram of HSV channels;
- **HSV_MS:** Histogram of HSV channels, the average and standard deviation of each channel;
- **HSV_RGB:** Histogram of HSV channels, Histogram of RGB channels, average and standard deviation of each channel of both color model;
- **LBP:** The descriptor resulted from LBP algorithm applied in a color image;
- **PFTAS:** The descriptor resulted from PFTAS algorithm applied in a color image;
- **GLCM:** The descriptor resulted from the metrics extracted of GLCM applied in a color image;
- **HED_LBP:** The descriptor resulted from LBP algorithm applied in a deconvoluted DAB image;
- **HED_PFTAS:** The descriptor resulted from PFTAS algorithm applied in a deconvoluted DAB image.
- **VGG16:** The descriptor resulted from the penultimate layer of VGG16 algorithm applied in a color image.
- **ResNet50:** The descriptor resulted from the penultimate layer of ResNet50 algorithm applied in a color image.

4.5 CLASSIFICATION

For both, image and patient level we experiment the following classifiers: SVM [10], KNN [12], MLP [37] and DT [5]. We evaluate the accuracy, precision and recall.

4.5.0.1 Image-level classification

We evaluate the features in the subset *feat_tr*, which is composed of patches that are representative of the WSI class. The best features are used to classify all the patches of a WSI, using *feat_tr* as training dataset. Thus, each patch of a WSI receives a label. However, instead of verifying it, we used it to generate a WSI's histogram of classes.

4.5.0.2 Patient-level classification

Although a WSI is scored in only one class, these slides may have patches from different classes. Therefore, we created a histogram of predicted patches' classes of each WSI and used this histogram as input for a classifier to determine the HER-2 score of the WSI.

5 RESULTS

5.1 PATCH-APPROACH

As described in Section 4.3, we choose a patch-approach in order to deal with processing problems due to WSIs are giga-images and cannot be entirely processed. Thus, a WSI is fully split into patches of size 250x250 pixels, creating a huge amount of small images. To avoid processing so many images, we proposed a two-steps algorithm for patches selection. Figure 5.1 and Figure 5.2 illustrate how much we reduce, in each WSI, the number of patches for both datasets Warwick and HistoBC-HER2.

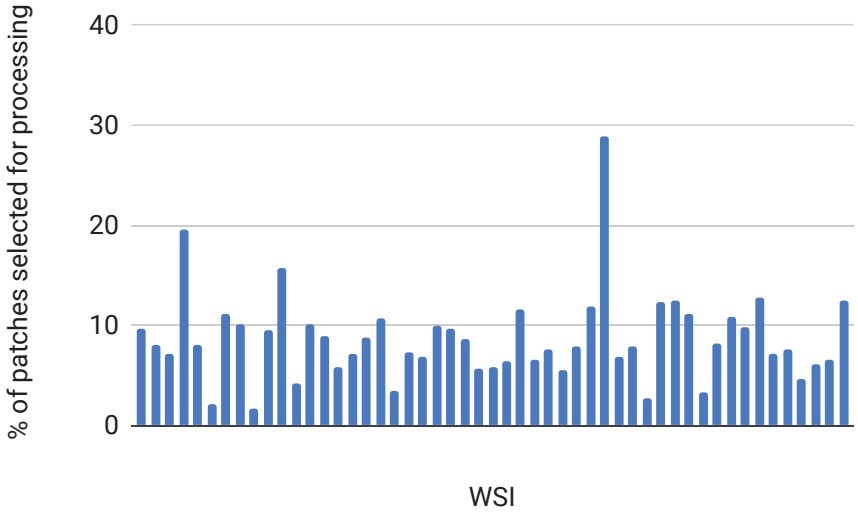


Figura 5.1: % of selected patches of each WSI for Warwick's dataset.
Source: The author.

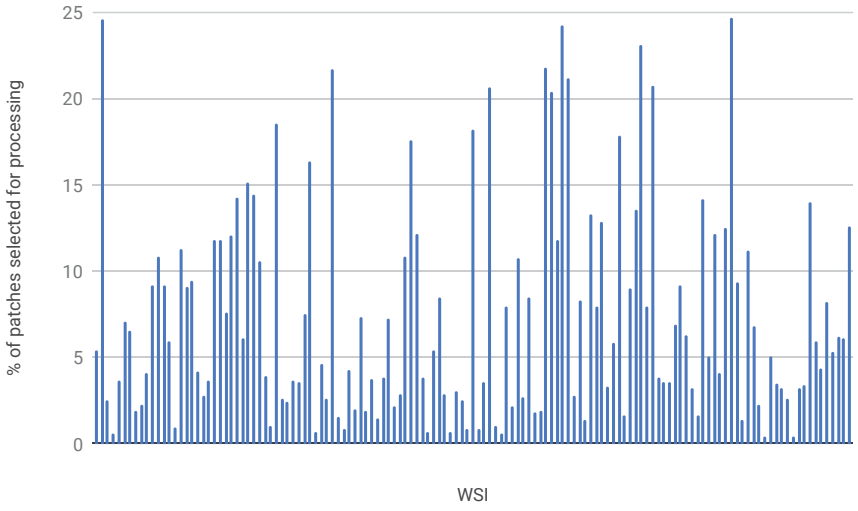


Figura 5.2: % of selected patches of each WSI for HistoBC-HER2 dataset.
Source: The author.

As observed in Figure 5.1 and Figure 5.2 the higher amount of resulted images in Warwick's dataset was around 30% of total patches extracted from a WSI. Meanwhile, in the HistoBC-HER2 dataset, the WSI with more patches selected resulted in 25% of its total patches. The average percent of select patches was 8.55% and 7.53%.

Although each WSI in each subset has a different reduction percentage, it does not means a fault in the preprocessing algorithm. Its performance must be evaluated according to the result of the classification since the important is the resulted images to be effective for the distinction of classes. This distinction in the percentage of patches reduction is due to the fact that the histological sections vary in size. Thus, the number of patches containing histological information also vary.

It is important to highlight that we not only presented an approach that allows the processing of WSIs but also reduces the processing in a more computationally expensive phase.

5.2 IMAGE-LEVEL

Since clinical decisions do not differentiate 0 and 1+ classes and only consider tests as negative (0/1+), borderline (2+) and positive (3+) [73], we have developed two approaches: with five (0, 1+, 2+, 3+ and noise) and four classes (negative, limit, positive and noise).

As we explained in Subsection 4.4 there is a subset created to evaluate algorithms for the classification of patches, labeling patches not according to the WSI class but with the tissue in the image.

Table 5.1 presents the result for the subset *feat_tr*. The explanation of the codification of the features is described in Subsection 4.4. As this subset requires a pathologist review, the same subset - from Warwicks's Dataset was used in the training phase of entire Warwicks and HistoBC-HER2 datasets.

Tabela 5.1: Accuracy on image level (in %) - Subset *feat_tr*. Bold shows the best results over a classifier and underlining shows the best results over all the features and classifiers.

	(0/1+), 2+, 3+ and NOISE				0, 1+, 2+, 3+ and NOISE			
	SVM	**KNN**	**MLP**	**DT**	**SVM**	**KNN**	**MLP**	**DT**
HSV	88.44	87.84	**90.27**	83.88	82.67	80.60	**86.26**	80.84
HSV_MS	88.62	**88.27**	90.14	85.55	82.77	**81.07**	85.54	**82.73**
HSV_RGB	88.36	86.90	89.68	**86.13**	82.55	79.64	84.91	81.38
LBP	58.37	51.21	56.46	50.67	50.80	41.89	49.21	39.68
PFTAS	79.87	73.16	76.60	68.05	69.77	65.14	69.12	59.07
GLCM	77.96	64.42	66.58	60.18	66.05	49.97	53.74	46.73
HED_LBP	76.16	70.52	72.65	64.28	73.18	66.24	68.54	59.87
HED_PFTAS	83.54	81.15	82.85	77.31	80.77	75.58	78.99	70.82
VGG16	87.74	82.16	86.63	73.10	82.50	75.52	81.30	67.50
ResNet50	**90.84**	87.84	89.70	83.51	**87.17**	81.00	85.25	76.97

Analyzing our results, the texture descriptors employed did not discriminate the evaluated color patches correctly. Even though their performance may be adversely affected by the interference generated during the conversion from DAB to gray levels, texture descriptors appear to be promising for deconvoluted images. Figure 5.3 shows some examples of this conversion.

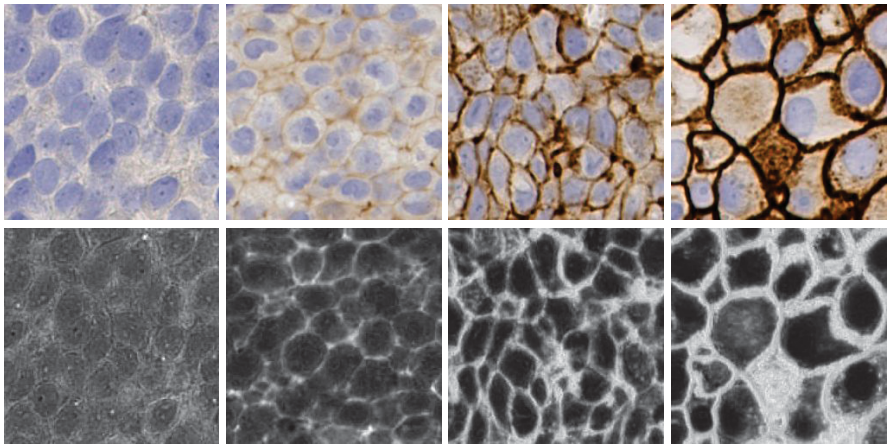


Figura 5.3: Example of deconvoluted D channel of DAB in each HER-2 class and converted to gray levels.

We obtained satisfactory results in both approaches, with four and five classes. As a WSI is composed of different classes of patches, these results seem to suggest representative patches are easier to classify than a WSI, which is heterogeneous.

We would like to highlight the best results were obtained from *ResNet50+SVM* for four classes and five classes. The second-best result was from *HSV+MLP* also for both groups of classes. These best results do not vary in a large percentage, which seems to suggest certain stability about the features.

As Table 5.1 shows in bold, the best features were those based on colors, thus they were used in image level, to distinguish patches and create the histogram to scoring HER-2. By using these histograms to predict the HER-2 score, we experimented SVM, KNN, MLP and DT at the patient level, the results of these experiments are present in the next section.

A challenge at this stage is to create a method robust to recognize the classes even with intra-class variations, as illustrated in Figure 5.4. As patches can have a different amount of tissue information and also color variation, due to ischemic time, tissue fixative and fixation time, tissue processing, the efficiency of epitope retrieval, selection of antibody or its clone and detection system [74].

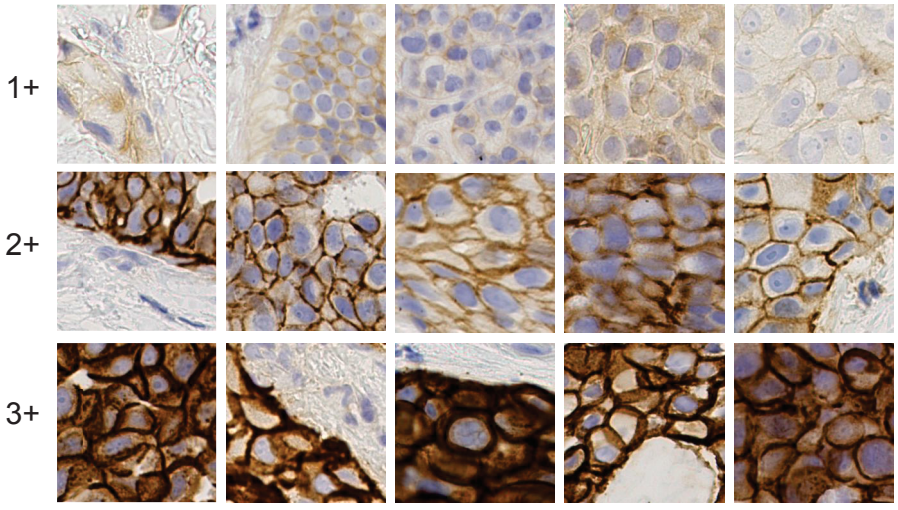


Figura 5.4: Examples of intra-class variations.

5.3 PATIENT-LEVEL

In the patient level, our algorithm uses as features the histogram of predicted patches' classes of each WSI generated in the image level. For example, *HSV+SVM* is the histogram of classes created by using HSV as a feature vector and SVM as a classifier to classify all the patches of a WSI in the image level.

5.3.1 Warwicks's Dataset

Regarding the performance at the patient level, Table 5.2 shows an overall increase in accuracy when classifying only with three classes (negative, borderline and positive), instead of all the four classes (0, 1+, 2+ and 3+). It is probably related to the similarity between 0 and 1+ classes that generate some confusion in trying to distinguish them.

Tabela 5.2: Accuracy on patient level - HER-2 scoring (in %). Bold shows the best results over a classifier and underlining shows the best results over all the features and classifiers.

	(0/1+), 2+ and 3+				0, 1+, 2+, 3+			
	SVM	**KNN**	**MLP**	**DT**	**SVM**	**KNN**	**MLP**	**DT**
HSV+SVM	80.39	74.51	68.63	60.78	66.67	62.75	56.86	60.78
HSV+KNN	80.39	78.43	64.71	66.67	70.59	60.78	60.78	62.75
HSV+MLP	76.47	82.35	72.55	72.55	66.67	70.59	64.71	68.63
HSV+DT	82.35	80.39	76.47	64.71	78.43	72.55	70.59	70.59
HSV_MS+SVM	64.71	72.55	62.75	50.98	66.67	50.98	60.78	58.82
HSV_MS+KNN	78.43	80.39	64.71	82.35	72.55	56.89	62.75	56.86
HSV_MS+MLP	82.35	86.27	74.51	72.55	72.55	60.78	54.90	64.71
HSV_MS+DT	80.39	**88.24**	74.51	76.47	68.63	68.63	68.63	58.82
HSV_RGB+SVM	74.51	80.39	54.90	68.63	70.59	64.71	56.86	68.63
HSV_RGB+KNN	82.35	76.47	76.47	68.63	70.59	62.75	56.86	58.82
HSV_RGB+MLP	80.39	82.35	66.67	68.63	70.59	74.51	56.86	66.67
HSV_RGB+DT	82.35	80.35	74.51	82.35	64.51	60.78	56.86	64.71
VGG16+SVM	86.27	86.27	84.31	82.35	66.67	76.47	74.51	**74.51**
VGG16+KNN	78.43	**88.24**	76.47	78.43	78.43	74.51	70.59	72.55
VGG16+MLP	86.27	**88.24**	**88.24**	76.47	**82.35**	**80.39**	**80.39**	72.55
VGG16+DT	70.59	78.43	72.55	68.63	60.78	60.79	60.78	49.02
ResNet50+SVM	84.31	86.27	86.27	76.47	66.67	70.59	66.67	58.82
ResNet50+KNN	78.43	**88.24**	80.39	74.51	74.51	78.43	76.47	72.55
ResNet50+MLP	**88.24**	88.24	86.27	**90.20**	70.59	74.51	74.51	72.55
ResNet50+DT	72.55	70.59	68.63	64.71	60.78	54.90	56.86	43.14

The best result at the patient level was obtained using *ResNet50+MLP* as a feature vector and DT as a classifier, which correctly predicted 90.20% of the WSIs, differentiating the negative (0/1+), borderline (2+) and positive (3+) classes. The second best score was 88.24% resulted from different feature vectors and classifiers. A problem is the reduced number of images, since we only have the GT of training subset in Warwick's dataset and it consists of 51 WSI, each image is a huge percentage in total. Since the best result in four-classes was 82.35%, the general decrease in accuracy when attempting to distinguish images in more classes is considerable.

Even though KNN is a simple classifier it performs well for the three-classes problem. It is possible to note that the KNN could distinguish well feature vectors of three-classes examples, as *HSV_MS+DT*, *VGG16+KNK*, *VGG16+MLP* and *ResNet50+KNN*.

The worst results in the three-classes approach are mainly by using *HSV_MS+SVM* feature vector. Although this combination performs well in the image level, such results suggest its confusions may impair the final score on the HER-2 score.

In order to improve the four-classes approach, we have tried to combine some of these 'histograms of classes' feature vectors. In a first attempt, we combined the best results in the patient level and in a second attempt, we combined the best results of the image level. However, we did not get any accuracy higher than 78.43% and therefore, the results of these experiments will not be reported.

Table 5.3 shows a confusion matrix of *ResNet50+MLP* and DT, the best accuracy obtained in three classes classification.

Tabela 5.3: Confusion Matrix of ResNet50+MLP classified by DT.

	Negative	**Borderline**	**Positive**
Negative	22	2	0
Borderline	1	12	1
Positive	1	0	12

In CADs focused on cancer treatment decision, it is important to evaluate precision and recall. By analyzing Table 5.3 it is possible to calculate both metric, and we present their results in Table 5.4.

Tabela 5.4: Precision and Recall of ResNet50+MLP classified by DT *per-class* (in %).

	Negative	**Borderline**	**Positive**
Precision	91.67	85.71	92.31
Recall	91.67	85.71	92.31

Precision and recall result into the same value, it indicates that errors are not prone to just one class. The 'borderline' class has the lowest recall, which means this method has more difficulty to identify scores 2+. Likewise, the 'borderline' class presented the lowest precision, suggesting our method failed when predicted an example as 2+. Additionally, these metrics demonstrate our method has a good performance in recognize 'negative' and 'positive' classes.

We observed all the 4 of 5 mistakes involved borderline (2+) class. We exemplified in Figure 5.5, a WSI with 2+ as GT that was scored as 3+. Figure 5.5-A shows a panoramic view of the tumor, it is notable that are many dark stains. However, as Figure 5.5-B and Figure 5.5-C show, some of these dark stains are scored as 3+ and others as 2+. Thus, it is a complex case to score. Another example is in Figure 5.6, in which both images have the same texture/morphology but the score was determined based on the proportion of this.

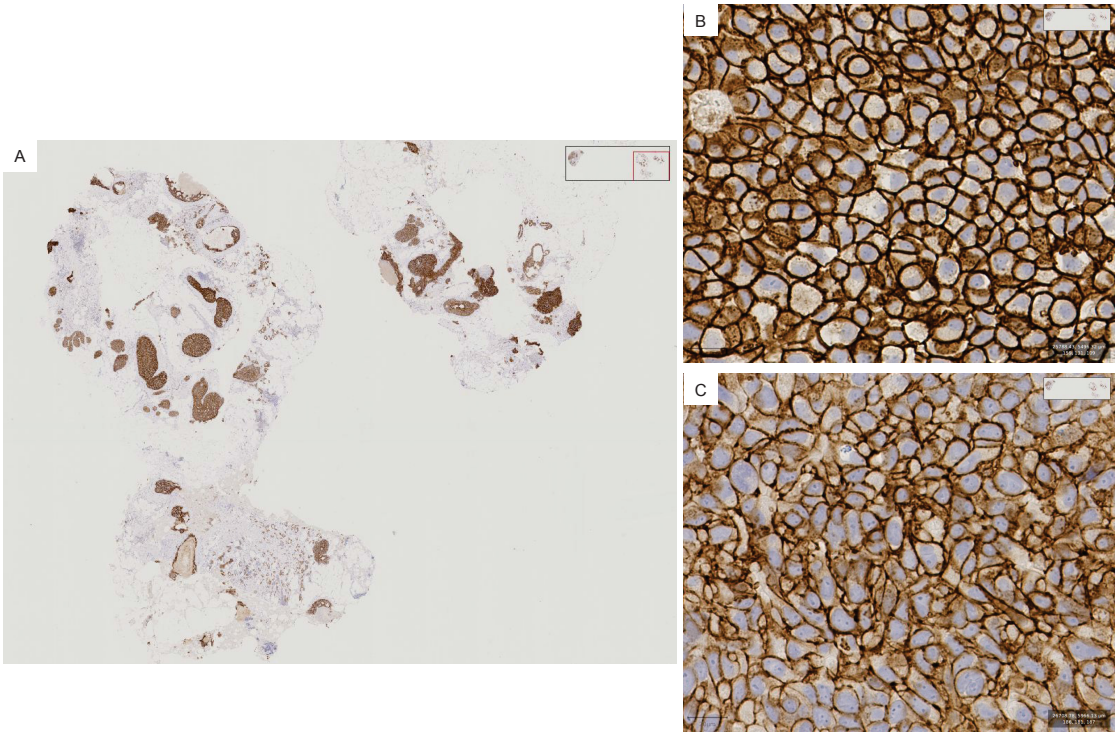


Figura 5.5: An illustration of a borderline WSI misclassified as positive. (A) Entire slide. (B) A 3+ tissue in the slide. (C) A 2+ tissue in the slide.

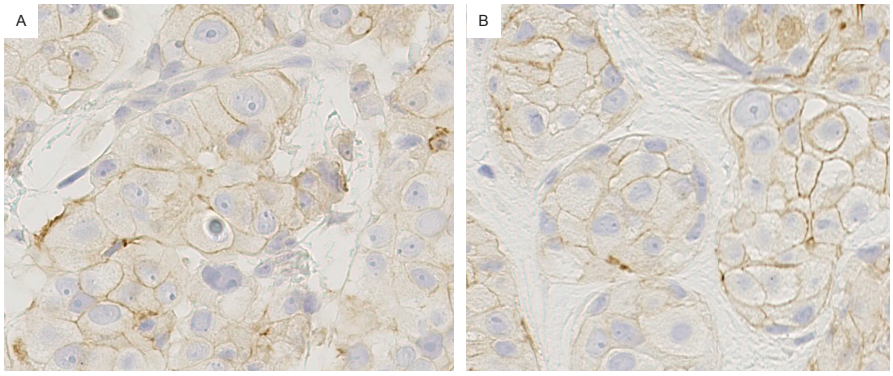


Figura 5.6: An illustration of two WSI misclassified. (A) A negative classified as borderline and (B) A borderline classified as negative.

5.3.2 HISTOBC-HER2 Dataset

HistoBC-HER2 is composed of 135 WSIs, more than 2 times Warwick size. Due to a lack of time, we decided to replicate in HistoBC-HER2 dataset the algorithms which performed better in Warwick's dataset.

5.3.2.1 Preprocessing results

As only the best algorithms are evaluated in HistoBC-HER2, another valuable analysis was performed, the effectiveness of the preprocessing algorithm. To evaluate the preprocessing potential, we first replicate the classification algorithms without patches selection explained in Subsection 4.3. We present this result in Table 5.5.

Tabela 5.5: Accuracy on patient level - HER-2 scoring (in %), in HistoBC-HER2 dataset without preprocessing. Bold shows the best results over a classifier and underlining shows the best results over all the features and classifiers.

	(0/1+), 2+ and 3+				0, 1+, 2+, 3+			
	SVM	**KNN**	**MLP**	**DT**	**SVM**	**KNN**	**MLP**	**DT**
VGG16+SVM	**62.30**	59.84	61.48	54.10	35.56	42.96	45.93	37.78
VGG16+KNN	61.16	61.16	60.33	59.50	45.93	44.44	42.22	42.22
VGG16+MLP	61.16	58.68	60.33	47.93	35.56	43.70	41.48	40.00
VGG16+DT	42.98	47.93	47.11	37.19	28.15	30.37	30.37	28.89
ResNet50+SVM	49.63	**63.70**	60.74	**66.67**	36.30	42.22	45.93	28.15
ResNet50+KNN	61.48	59.26	**63.70**	54.81	**47.41**	43.70	**46.67**	41.48
ResNet50+MLP	57.04	57.78	61.48	59.26	26.67	**47.41**	39.26	**45.93**
ResNet50+DT	47.41	42.22	47.41	33.33	27.41	30.37	37.78	24.44

Afterward, we entirely reproduce the proposed algorithm. This outcome is shown in Table 5.6. According to Table 4.2, the P1 and P2 reviews are not well balanced among classes. Thus, results presented here are based on a classifier trained with Record annotations and later compared with P1 and P2.

Tabela 5.6: Accuracy on patient level - HER-2 scoring (in %), in HistoBC-HER2 dataset with preprocessing. Bold shows the best results over a classifier and underlining shows the best results over all the features and classifiers.

	(0/1+), 2+ and 3+				0, 1+, 2+, 3+			
	SVM	**KNN**	**MLP**	**DT**	**SVM**	**KNN**	**MLP**	**DT**
VGG16+SVM	71.85	74.07	75.56	74.81	57.78	54.81	47.41	49.63
VGG16+KNN	71.85	75.56	64.44	65.18	52.59	54.81	53.33	**57.04**
VGG16+MLP	68.15	71.11	63.70	76.30	54.81	55.56	50.37	54.07
VGG16+DT	70.37	70.37	77.04	64.44	53.33	57.78	51.11	47.41
ResNet50+SVM	76.30	63.70	**77.78**	64.44	52.59	**60.00**	**56.30**	51.11
ResNet50+KNN	**79.26**	**76.30**	74.07	77.78	**60.00**	51.11	54.07	56.30
ResNet50+MLP	78.52	75.56	68.89	**78.52**	55.56	56.30	53.33	51.11
ResNet50+DT	72.59	63.70	77.04	77.78	56.30	55.56	52.59	51.85

As Tables 5.5 and 5.6 are showing, the best result without preprocessing was 66.67% and 79.26% with preprocessing, both for 3 classes problem. Although the best classifier was different, the histogram result from *ResNet50* features shows better performances. Comparing these results, we notice an increase of 12,59% from one best result to another. This increase in accuracy may be attributable to the preprocessing algorithm. This step should select the most relevant patches to be analyzed, resulting in the processing of images with information of greater distinction between classes, reducing the classification confusion.

5.3.2.2 General Results

General results are presented in Table 5.6. Analyzing the best result overall HistoBC-HER2 dataset, it was obtained using *ResNet50+KNN* as a feature vector and SVM as a classifier, which correctly predicted 79.26% exams distinguish them into 3 classes (negative, borderline and positive).

Table 5.7 shows a confusion matrix of the best result. As mentioned before, for CADs precision and recall metrics are important. Table 5.7 present these values obtained with *ResNet50+KNN* and SVM.

Tabela 5.7: Confusion Matrix of ResNet50+KNN classified by SVM.

	Negative	**Borderline**	**Positive**
Negative	59	5	2
Borderline	15	18	2
Positive	1	3	30

Tabela 5.8: Precision and Recall of ResNet50+KNN classified by SVM *per-class* (in %).

	Negative	**Borderline**	**Positive**
Precision	78.67	69.23	88.24
Recall	89.39	51.43	88.24

The borderline class has the lowest metrics, which means the algorithm makes confusion involving it. Although, the method is good to recognize negative class and even more to positive.

The result presented only three mistakes between positive and negative classes. One wrongly classified a positive as negative and the other two the opposite. Figure 5.7 illustrates the first case. Our analysis, as non-experts, is that the image suggests a real negative case, as reviewed by pathologist 1. Figure 5.8 and 5.9 shows that both negative images have some positive marks for HER-2, but not enough as recommended by the UK guideline in Table 2.1, to be positive, so they are real mistakes.

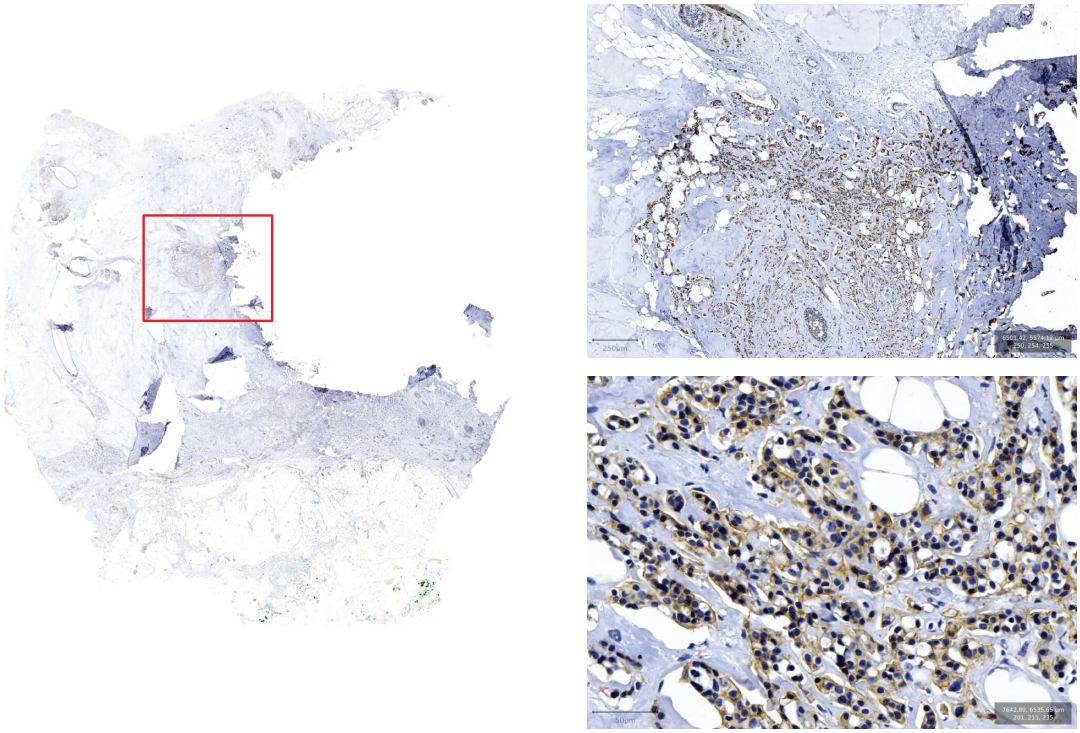


Figura 5.7: 062 - A positive image classified as negative.
Source: The Author.

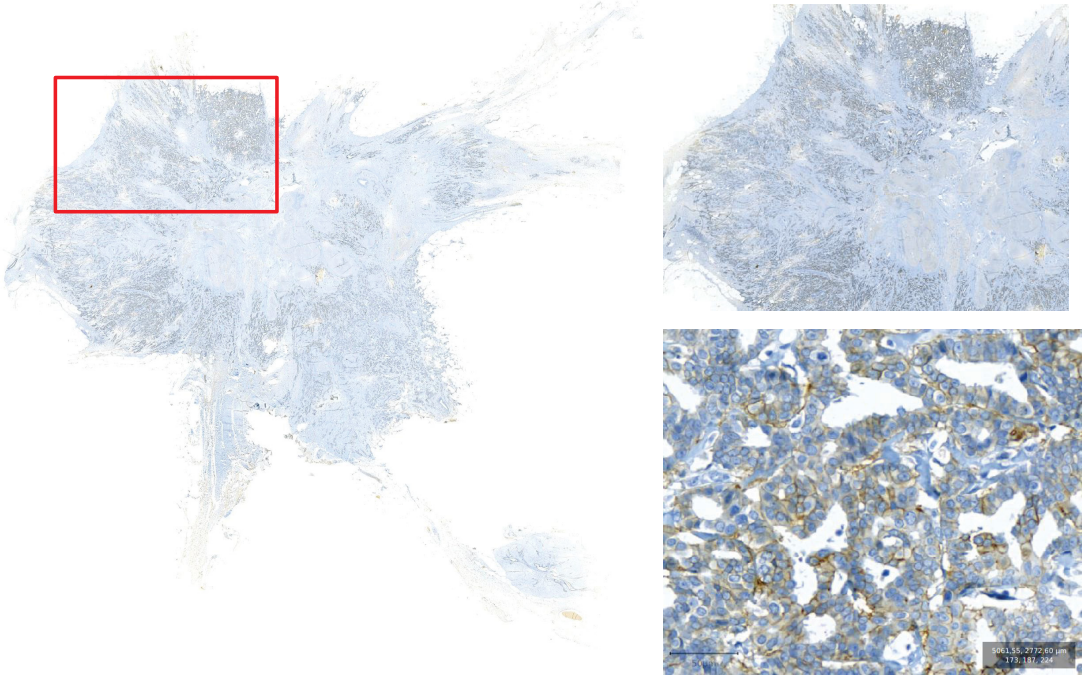


Figura 5.8: WSI 385 - A negative image classified as positive.
Source: The Author.

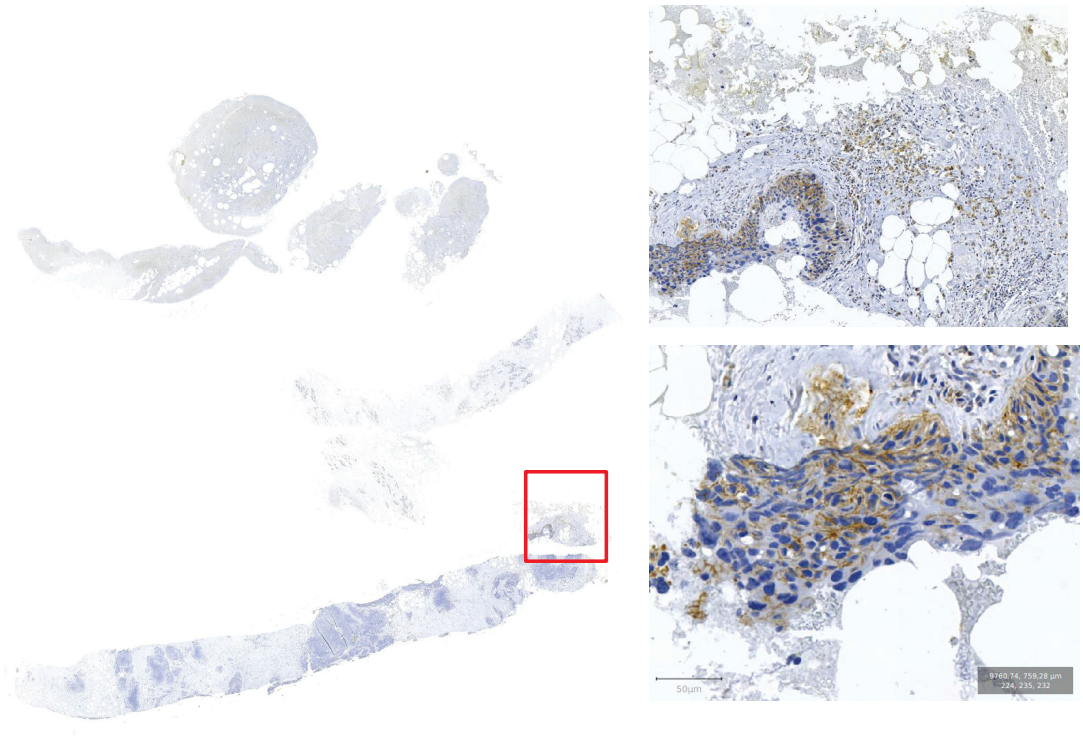


Figura 5.9: WSI 477 - A negative image classified as positive.
Source: The Author.

5.3.2.3 Easy-Medium-Hard Images Results

As described in Section 4.2 each WSI in HistoBC-HER2 dataset was consider as easy, medium or hard to score, according to the number of agreement among all reviews. We present below the accuracy for each degree of difficulty.

- **Easy:** 80.85%
- **Medium:** 81.16%
- **Hard:** 68.42%

Since we have these disagreements among record and pathologists analysis, we decided to evaluate the algorithm only based on the record, due to classes are balanced. Subsequently to compare the mistakes in the classification with the pathologist's reviews. Tables 5.9, 5.10 and 5.11 present this comparative. We write in bold the exams in which mistakes have resulted from the classifier and still, they have an agreement with a pathologist. We use N to represent negative for HER-2, B for borderline and P for positive.

Table 5.9 shows that *ResNet50+KNN* classified with SVM confused classes from 9 of 47 exams.

Tabela 5.9: HistoBC-HER2 dataset - Easy images Results

Easy images														
Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result
001	P	P	P	P	224	B	B	B	B	513	P	P	P	B
020	N	N	N	N	233	P	P	P	P	518	P	P	P	B
056	P	P	P	P	238	P	P	P	P	519	B	B	B	N
058	P	P	P	P	239	P	P	P	P	541	P	P	P	P
095	P	P	P	P	251	B	B	B	N	631	P	P	P	P
097	P	P	P	P	276	P	P	P	P	649	B	B	B	N
103	P	P	P	P	304	B	B	B	B	661	P	P	P	P
121	B	B	B	N	334	P	P	P	P	662	P	P	P	P
136	P	P	P	P	338	P	P	P	P	667	N	N	N	N
141	B	B	B	N	358	P	P	P	P	670	P	P	P	P
149	P	P	P	P	398	P	P	P	P	679	B	B	B	B
169	P	P	P	P	447	P	P	P	P	695	B	B	B	N
178	P	P	P	P	470	P	P	P	P	748	B	B	B	B
203	P	P	P	P	476	N	N	N	N	864	B	B	B	B
207	P	P	P	P	479	P	P	P	P	865	B	B	B	N
					491	N	N	N	N	887	B	B	B	B

Table 5.10 give us the amount of 13 misclassified exams from a total of 69. However, 10 of 13 have an agreement with some pathologist.

Tabela 5.10: HistoBC-HER2 dataset - Medium images Results

Medium images														
Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result
003	N	N	N	N	202	N	N	N	N	442	N	B	B	N
008	N	N	N	N	216	N	B	B	N	448	B	N	B	B
015	N	N	N	N	225	N	N	N	N	451	N	N	N	N
016	N	N	N	N	232	N	N	B	N	457	N	N	N	N
018	N	N	B	N	266	N	N	B	N	458	B	N	B	B
024	B	**N**	B	**N**	307	B	**N**	B	**N**	460	B	N	B	B
043	N	N	**B**	**B**	329	N	N	N	N	477	N	N	N	P
049	N	N	B	N	331	N	N	N	N	478	N	N	B	N
054	N	N	N	N	335	N	N	N	N	490	N	N	N	N
063	N	N	N	N	349	B	**P**	**P**	**P**	498	N	N	**B**	**B**
073	N	N	N	N	355	N	B	B	N	499	N	N	B	N
074	N	N	N	N	360	N	N	N	N	520	B	N	B	B
077	N	N	N	N	363	N	N	B	N	534	N	N	N	N
102	B	**N**	B	**N**	366	N	N	N	N	588	B	N	B	B
161	N	N	N	N	377	B	N	B	B	625	N	N	N	N
163	B	N	B	B	378	B	N	B	B	630	B	P	B	B
166	N	N	N	N	380	N	N	N	N	693	N	N	B	N
168	P	**B**	**B**	**B**	382	N	N	N	N	699	B	N	B	P
170	N	N	N	N	385	N	B	B	P	785	B	N	B	B
171	B	**N**	B	**N**	386	N	N	N	N	823	B	**N**	B	**N**
182	N	N	N	N	391	N	N	N	N	859	B	**N**	**N**	**N**
191	P	B	P	P	400	N	N	B	N	875	B	N	B	B
195	N	N	N	N	441	N	B	B	N	899	B	N	B	B

Table 5.11 presents the result of hard images, where 6 of 19 exams were wrongly classified. Even though all of them have one agreement, we can not consider it as a correct prediction. Due to considering only 3 classes, and, hard images the ones with disagreements

among the 3 pathologists, each analysis should be from one class, then they will consequently agree with some other analysis.

Tabela 5.11: HistoBC-HER2 dataset - Hard images Results

Hard images														
Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result	Exam	Rec	P1	P2	Result
017	N	P	N	N	184	N	N	B	N	379	N	N	B	N
022	N	N	**B**	**B**	214	N	N	B	N	444	N	N	B	N
051	N	N	**B**	**B**	218	N	N	B	N	459	N	N	B	N
053	B	**N**	P	**N**	220	N	N	**B**	**B**	480	N	N	B	N
062	P	**N**	B	**N**	273	N	N	B	N	638	P	N	N	P
096	B	**N**	**N**	**N**	325	N	P	B	N	668	N	N	B	N
										707	N	N	B	N

Medium images had 13 mistakes, but 10 agreements with at least one pathologist. From these 10 agreements, 3 had an agreement between the two pathologists.

The WSI number 168 has HER-2 Positive in the Record, but the two pathologists analyzed as Borderline. The algorithm agrees with the pathologists. Figure 5.10 shows the slide has few marks and the borders are difficult to classify. This is a difficult case, which the algorithm could assistant with a second opinion.

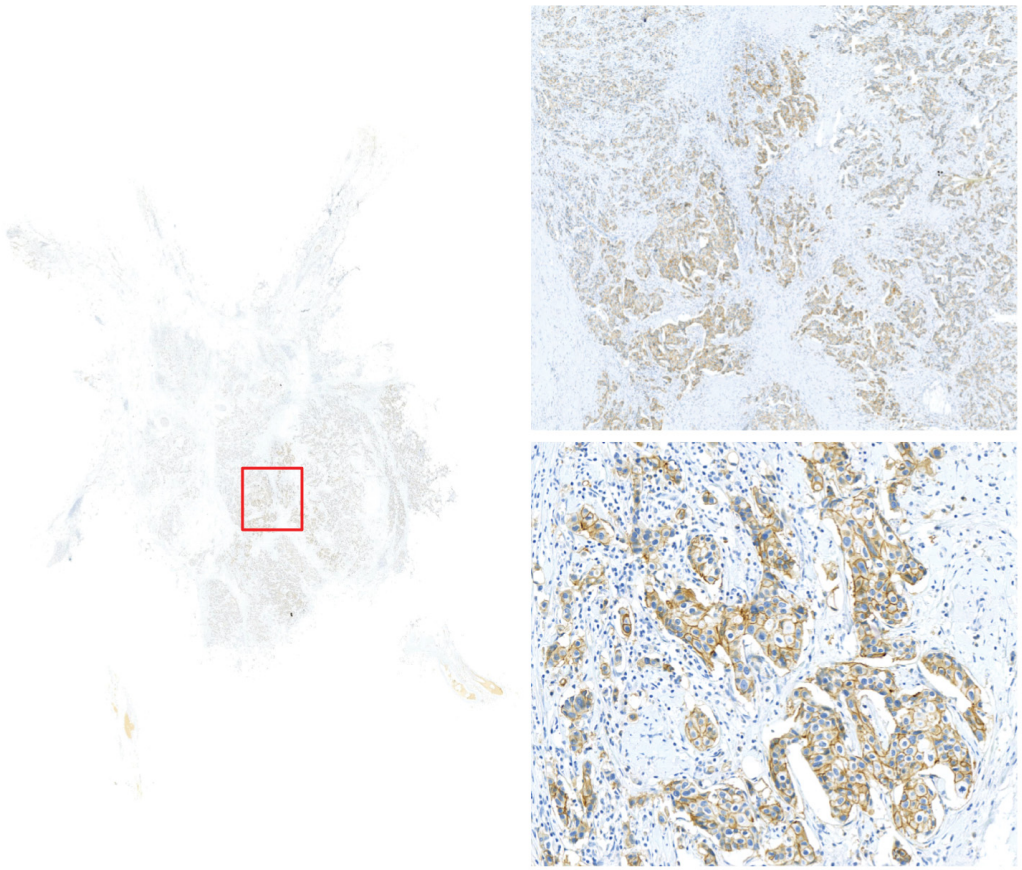


Figura 5.10: WSI 168
Source: The Author.

Other cases that have an agreement between pathologists and the algorithm, but disagreement with the patient record are WSI 349 and 859.

Figure 5.11 illustrates some views from WSI 349. Pathologists and the algorithm classified this as HER-2 Positive, but in the patient records it is Borderline. Although the slide has enough markup, not all are continuous and intense. The marks range from possible marking to 1+ and 2+, making it difficult to get an accurate result

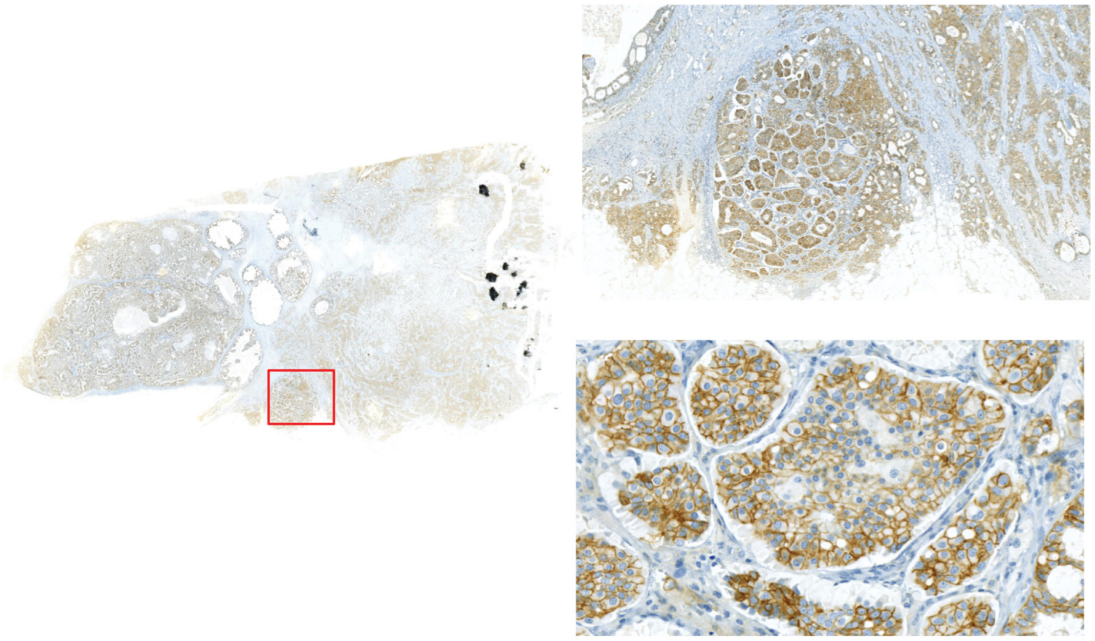


Figura 5.11: WSI 349
Source: The Author.

Likewise, Figure 5.12 illustrates some views from WSI 859. Only the patient record takes this exam as Borderline. Pathologists and algorithm agreed with an HER-2 negative case. As we can see, there are few membranes marked-up and with low intensity.

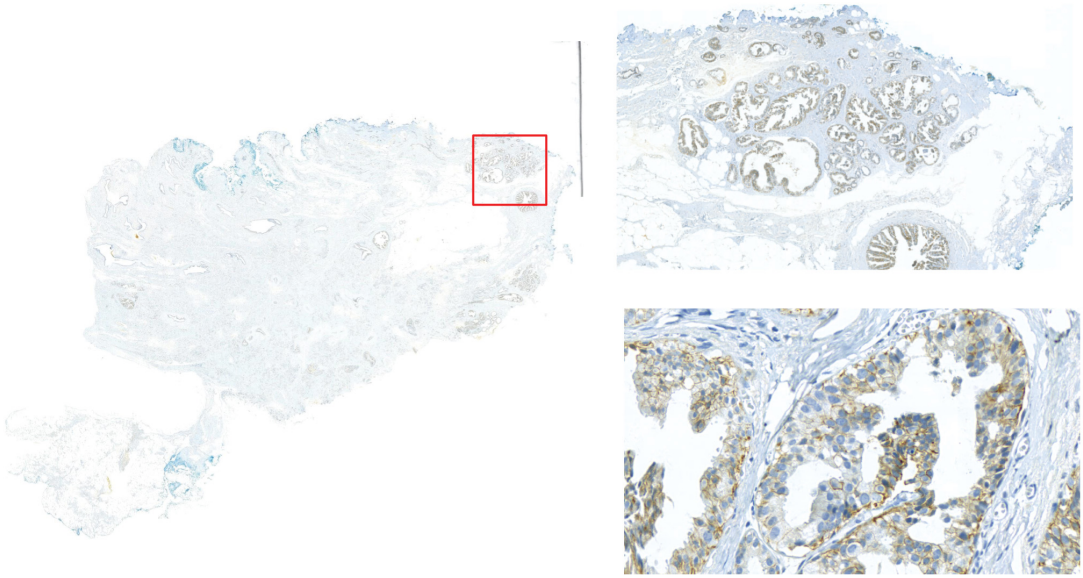


Figura 5.12: WSI 859
Source: The Author.

Analysing Tables 5.9, 5.10 and 5.11 we might consider 10 of 27 , firstly considered as mistakes, as correct. Only the 10 of 13 cases of medium-difficult images which have some other agreement. Resulting in an agreement value of 86.66%.

5.4 FINAL REMARKS

Mistakes between 2+ and other scores are very common, thus a FISH test is required to confirm HER-2 positivity in 2+ slides. Despite some borderline confusions, our results are still very promising and might assist pathologists as a second opinion.

The best accuracies in Warwick's dataset were 90.20% for three class-problem and 82.23% for the four-class problem. This result for negative, borderline and negative approach has resulted from ResNet50+MLP in the image level and DT at the patient level. The result for 0, 1, 2+ and 3+ approach, was obtained by using VGG16+MLP in image level and SVM at the patient level.

Whereas, in the HistoBC-HER2 dataset the best accuracies were 79.26% and 60.00% for three and four classes approach, respectively. The three-class approach was better differentiated with a combination of ResNet50+KNN in the image level and SVM at the patient level. And the arrangement of ResNet50+KNN at the image level with SVM at the patient level. As also ResNet50+SVM and KNN presented the best result in the four-class approach.

We noticed that SVM and KNN performed well in patient level when using ResNet50 as descriptor at the image level.

Moreover, the relevance of the preprocessing can be noticed when comparing classification results with and without this step. The best accuracy of HistoBC-HER2 dataset was 79.26%, obtained by ResNet50+KNN and *SVM*. The same algorithm without preprocessing presented an accuracy of 61.48%.

6 CONCLUSION

The objective of this work was to provide a technique able to score HER-2 in WSI, focusing on avoiding segmentation and manual intervention, since both can introduce a subjective criterion to the classification.

As described in literature review, most classical approaches include segmentation, which is known to introduce errors in the next steps. Their concordance was around 85%, being increased by using deep learning techniques. Nonetheless, our approach achieved more than 85% accuracy, avoiding explicit segmentation and extraction of structure properties such as cell nuclei, membrane, size and shape of these. We concluded in the literature review the lack of available datasets that allow the development of other works. Thus, this work provides a new WSI dataset, named HistoBC-HER2, IHC tests were conducted for HER-2 in cases of BC.

In order to score HER-2 avoiding segmentation, we proposed a methodology divided into the image and patient level, each one of them was individually evaluated. The evaluation of the image level required a creation of a subset for training patches classes, which we named *feat_tr*. At this level ten features vectors and four classic classifiers were employed. Among these features vectors there are color and textures descriptors and others that were extracted from CNNs.

Moreover, we adopted two approaches for classes determination - clinical decision and HER-2 scoring. For the clinical decision the classes are 'negative', 'borderline' and 'positive'. The HER-2 scoring differs 0, 1+, 2+ and 3+ classes.

Promising results were obtained in Warwick's dataset, 90.20% of accuracy. An obvious limitation of this result is the number of images experimented. For this reason, we introduced a new dataset, HistoBC-HER2, where our algorithms could be evaluated more robustly. The results of this new dataset would seem to suggest an effectiveness preprocessing algorithm. As we evaluated our method with and without preprocessing, and the best accuracy were 66.67% and 79.26%, respectively.

Our method avoids segmentation and do not need manual intervention, different of several works reviewed. In Table 6.1 we compared our method with others described before. It is hard to compare the results since images and protocols are different for each work. The HER2NET proposed by [59] had a better accuracy than ours. Although both works have used the same dataset, partitions for training and test were different. Also, HER2NET depends on manual intervention for ROI selection and includes a segmentation step.

Tabela 6.1: Comparison with related works - IHC Images using Classical Image Processing

	Manual Intervention	**Segmentation**	**Remarks**
[42] (2017)	Yes	No	88.46% accuracy
[69] (2013)	No	No	83% accuracy
[32] (2012)	Yes	No	0.72 Kendall τb
[73] (2012)	Yes	Yes	$k_w = 80$
[41] (2009)	Yes	Yes	81% accuracy
[21] (2008)	No	Yes	64% accuracy
Proposed work	**No**	**No**	**90.20% accuracy**

Since promising results were achieved, we can affirm that a further contribution of this work is the proposed pre-processing step, which reduces the number of patches to be processed. Besides, it is fully automated and can easily works in simple desktop computers. Thus, findings presented in this work support the idea of cheap techniques to help in pathologists routine.

Furthermore, we propose in future work to compare different sizes of patches and use all a CNN including the classification task. Additionally, some improvements in the *feat_tr* subset are required, such as a review and an increase in the number of images, by a expert selection or by data augmentation.

REFERÊNCIAS

[1] T. Ahonen et al. Rotation invariant image description with local binary pattern histogram fourier features. In *Image Analysis*, pages 61–70, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] P. E. Aktan, G. Hatipoğlu, and N. Arica. Risk classification for breast cancer diagnosis using her2 testing. In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 2133–2136, May 2016.

[3] Yoshua Bengio and Aaron Courville. Deep learning of representations. *Handbook on Neural Information Processing*, 49, 2015.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] L. Breiman. *Classification and Regression Trees*. Routledge, 1984.

[6] C. Brito, M. C. Portela, and M. T. L. D. Vasconcellos. Assistência oncológica pelo sus a mulheres com câncer de mama no estado do rio de janeiro. *Revista de Saúde Pública*, 39:874–881, 2005.

[7] A. Brügmann et al. Digital image analysis of membrane connectivity is a robust measure of her2 immunostains. *Breast Cancer Research and Treatment*, 132(1):41–49, 2012.

[8] N. E. Buckley et al. Quantification of her2 heterogeneity in breast cancer–implications for identification of sub-dominant clones for personalised treatment. *Scientific reports*, 6:23383, 2016.

[9] L. P. Coelho et al. Structured literature image finder: extracting information from text and images in biomedical literature. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2–3), 2010.

[10] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[11] Ministério da Saúde. Portaria nº 29, de 2 de agosto de 2017. Visited em 03/05/2018.

[12] B. V. Dasarathy. *Nearest Neighbor (NN) Norms NN pattern Classification Techniques*. IEEE Computer society press, 1991.

[13] C. DeSantis et al. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.

[14] L. Dobson et al. Image analysis as an adjunct to manual her-2 immunohistochemical review: A diagnostic tool to standardize interpretation. *Histopathology*, 57(1):27–38, 2010.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

[16] J. Ferlay et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015.

[17] M. W. Gardner and S. R. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[18] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson Prentice Hall, 2008.

[19] A. Goode et al. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[20] Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.

[21] B. H. Hall et al. Computer-assisted assessment of the human epidermal growth factor receptor 2 immunohistochemical assay in imaged histologic sections using a membrane isolation algorithm and quantitative analysis of positive controls. *BMC Medical Imaging*, 8:1–13, 2008.

[22] N. A. Hamilton et al. Fast automated cell phenotype image classification. *BMC Bioinformatics*, 8, 2007.

[23] P. W. Hamilton et al. Digital pathology and image analysis in tissue biomarker research. *Methods*, 70(1):59–73, 2014.

[24] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[25] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 3(6):610–621, 1973.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] H. Holten-Rossing et al. Optimizing her2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Research and Treatment*, 152(2):367–375, 2015.

[28] L. Hou et al. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

[29] INCA. Estimativa 2014 – incidência de câncer no brasil. `http://www.inca.gov.br/wcm/dncc/2013/apresentacao-estimativa-2014.pdf`, 2014. Visited em 08/11/2015.

[30] INCA. Estimativa 2018 – incidência de câncer no brasil. `http://www.inca.gov.br/estimativa/2018/index.asp`, 2018. Visited em 03/05/2018.

[31] A. S. Joshi et al. Semi-automated imaging system to quantitate her-2/neu membrane receptor immunoreactivity in human breast cancer. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 71(5):273–285, 2007.

[32] B. Keller, W. Chen, and M. A. Gavrielides. Quantitative assessment and classification of tissue-based biomarker expression with color content analysis. *Archives of Pathology and Laboratory Medicine*, 136(5):539–550, 2012.

[33] T. Koopman et al. Digital image analysis of her2 immunohistochemistry in gastric- and oesophageal adenocarcinoma: a validation study on biopsies and surgical specimens. *Histopathology*, 72(2):191–200, 2018.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[35] V. Kumar, A. K. Abbas, and J. C. Aster. *Robbins Basic Pathology*. Elsevier Health Sciences, 2013.

[36] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[37] Y. A. LeCun et al. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 9–50. Springer-Verlag, 1998.

[38] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256. IEEE, 2010.

[39] R. S. Ledley, M. Buas, and T. J. Golab. Fundamentals of true-color image processing. In *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, volume 1, pages 791–795, Jun 1990.

[40] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013.

[41] H. Masmoudi et al. Automated quantitative assessment of her-2/neu immunohistochemical expression in breast cancer. *IEEE Transactions on Medical Imaging*, 28(6):916–925, 2009.

[42] R. Mukundan. A robust algorithm for automated her2 scoring in breast cancer histology slides using characteristic curves. *Communications in Computer and Information Science*, 723:386–397, 2017.

[43] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, Jul 2002.

[44] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[45] Breast Cancer ORG. Her2 status. Visited em 10/06/2019.

[46] WHO. World Health Organization. Global cancer observatory, 2014.

[47] F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[48] T. Pitkäaho et al. Classifying her2 breast cancer cell samples using deep learning. In *Irish Machine Vision and Image Processing Conference Proceedings 2016*, pages 78–85, Galway - Ireland, August 2016.

[49] Moacir Antonelli Ponti, Leonardo Sampaio Ferraz Ribeiro, Tiago Santana Nazare, Tu Bui, and John Collomosse. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)*, pages 17–41. IEEE, 2017.

[50] T. Qaiser et al. Her2 challenge contest: a detailed assessment of automated her2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*, 72(2):227–238, 2018.

[51] F. Raimondo and M. A. Gavrielides. Automated evaluation of her-2/neu status in breast tissue from fluorescent in situ hybridization images. *IEEE Trans. on Image Processing*, 14(9):1288–1299, 2005.

[52] E. A. Rakha et al. Updated uk recommendations for her2 assessment in breast cancer. *Journal of Clinical Pathology*, 2014.

[53] S. Razavi, G. Hatipoğlu, and H. Yalçın. Automatically diagnosing her2 amplification status for breast cancer patients using large fish images. In *Signal Processing and Communications Applications Conference (SIU), 2017 25th*, pages 1–4. IEEE, May 2017.

[54] L. Robert. *Learning Data Mining with Python*. Packt Publishing, 2015.

[55] E. Rodner, M. Simon, and J. Denzler. Deep bilinear features for her2 scoring in digital pathology. *Current Directions in Biomedical Engineering*, 3(2):811–814, 2017.

[56] L. Rokach and O. Maimon. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.

[57] A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, 2001.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[59] M. Saha and C. Chakraborty. Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, 7149(c):1–1, 2018.

[60] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[62] I. Skaland et al. Comparing subjective and digital image analysis her2/neu expression scores with conventional and modified fish scores in breast cancer. *Journal of Clinical Pathology*, 61(1):68–71, 2008.

[63] D. J. Slamon et al. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *New England Journal of Medicine*, 344(11):783–792, 2001.

[64] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[65] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE, 2016.

[66] B. Stenkvist et al. Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations. *Cancer research*, 38(12):4688–4697, 1978.

[67] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[68] J. Słodkowska et al. Study on breast carcinoma her2/neu and hormonal receptors status assessed by automated images analysis systems: Acis iii (dako) and scanscope (aperio). *Folia Histochemica et Cytobiologica*, 48(1):19–25, 2010.

[69] M. Tabakov. Using fuzzy sugeno integral as an aggregation operator of ensemble of fuzzy decision trees in the recognition of her2 breast cancer histopathology images. In *Computer Medical Applications (ICCMA), 2013 International Conference on*, pages 1–6. IEEE, 2013.

[70] P. N. Tan et al. *Introduction to Data Mining*. Addison-Wesley, 2005.

[71] Z. Theodosiou et al. Evaluation of fish image analysis system on assessing her2 amplification in breast carcinoma cases. *Breast*, 17(1):80–84, 2008.

[72] M. Tkalcic and J. F. Tasic. Colour spaces: perceptual, historical and applicational background. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, volume 1, pages 304–308, Sept 2003.

[73] V. J. Tuominen, T. T. Tolonen, and J. Isola. Immunomembrane: A publicly available web application for digital image analysis of her2 immunohistochemistry. *Histopathology*, 60(5):758–767, 2012.

[74] A. R. Vahadane. *A Few Algorithms for Histopathological Images in Computational Pathology*. PhD thesis, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India, April 2017.

[75] S. Van der Walt et al. Scikit-image: image processing in python. *PeerJ*, 2:e453, 6 2014.

[76] M. E. Vandenberghe et al. Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer. *Scientific Reports*, 7(March):45938, 2017.

[77] G. Viale et al. Assessment of her2 amplification status in breast cancer using a new automated her2 iqfish pharmdx™ (dako omnis) assay. *Pathology Research and Practice*, 212(8):735–742, 2016.

[78] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *European Symposium on Artificial Neural Networks*, Bruges - Bélgica, Abril 1999.

[79] C. A. Yamamoto. *Revisão integrativa sobre monitoramento de Programas de Controle do Câncer de Mama*. PhD thesis, Fundação Oswaldo Cruz. Escola Nacional de Saúde Pública Sergio Arouca, 2018.

[80] H. Yaziji et al. Her-2 testing in breast cancer using parallel tissue-based methods. *Jama*, 291(16):1972–1977, 2004.